NOT TO BE FURTHER COPIED OR REPRODUCED

(* y VS x *)

10

Tilts Are There To Be Taken Away

		Exh.		Page
Introduction 10-1				
untilted, at 10-2		1		10-3
tilt, at 10-2		2		10-4
- a line for three points 10-5a			A	
			B	
		3		10-13/16
10A.	A simple approach to tilt assessment 10-6	4		10-22
	- use of three portions 10-6		C	
	- rule of thumb 10-7	5		10-24/25
	- choice of origin 10-7	6		10-27
	- using a trial tilt 10-8		D	
	- taking a 2nd step? 10-9	7		10-29/29a
	- a weather eye for bend 10-10	8		10-31
	- finishing off 10-10	9		10-32/33a
		10		10-34
		11		10-36
10B.	Untilting with means 10-12	12	E	10-38/43
	- speed and cost of multiplication 10-12		F	
	- step-by-step untilting 10-19	13		10-45
	- a snare 10-20	14		10-47/48
	- looking at the residuals 10-21	15		10-50
			G	
10C.	Another example, value of imports 10-23		H	
			K	
		16		10-57/60
10D.	That middle-mean flavor 10-28	17		10-61
		18		10-62
10E.	A larger example -- 49 states 10-37	19		10-64
	- marriages and m.v. deaths 10-37		L	
		20		10-67/68
10F.	Looking at the residuals 10-44	21		10-70
	- looking individually 10-46	22		10-71
		23		10-72/73
10G.	Some remarks (background) 10-51	24		10-74
		25	M	10-76
	**********		P	
10H.	How far have we come?			
	**********			
10K.	A functional example (optional but instructive) 10-56			
	**********			
10L.	When computing is no pain (optional background or optional with computer 10-65			
	- short-cutting? 10-69			
	**********			
10M.	We could have skipped states (optional, uses chapter 6) 10-75			
	**********			
10P.	Problems for chapter 10 10-77			

10

Tilts Are There To Be Taken Away

Clearing away big -- or obvious -- effects so as to see smaller -- or camouflaged -- ones is a major task of exploratory data analysis. Indeed one can argue that almost every one of its techniques does this in one way or another. Residuals have had the fit subtracted so we can see what remains. When the typical values of two groups are compared, one with another, we can see small differences even when all typical values are very large. And so on.

The last step we took in chapter 8, as part of a general program of improving plots, was to flatten them. Operating graphically, we could do this quite well by eye. Operating numerically, as we also wish to do (not because it is harder but because complex cases tend to escape graphical treatment), we need simple and effective techniques for choosing fits. Their description and use begins in this chapter.

When we operate graphically, we can try to straighten the picture out before flattening it. Numerical approaches are almost never as flexible as graphical ones. Accordingly we will want to have a procedure that flattens the data out whether or not it is straight -- whether or not it has been straightened. Here "flattening" must refer to the general run of the data, and not to its detailed behavior. Something

that looks like a plausible set of residuals from a straight line is "flattened", though it may not be flat. Exhibit 18 of chapter 8 is an example.

Exhibit 1 shows a rather extreme example of curvature and exhibit 2 the results of "flattening" exhibit 1. Clearly exhibit 2 is far from being flat, so "flattening" is a poor term. Equally, it is poor use of words to talk about "the slope" of exhibit 1, since there is a very small slope toward the left of the plot and a very large slope to the right.

Tilt is a short word, and conveys an appropriate feeling. We shall say that exhibit 2 is

untilted

and that the slope of the line whose subtraction converts exhibit 1 into exhibit 2 -- thus clearing away an important part of what is going on so that we can see better what remains -- is the

tilt

of exhibit 1. (If you regard this kind of a definition as backhanded, read section 11D.)

Thus, we take the question "How much is y tilted against x ?" to mean: "What value of b leaves $y - bx$ apparently untilted?" . Equally, we shall take the b thus found as our assessment of tilt, and the values of $y - bx$ as what we have been able to do about freeing y of tilt against x.

exhibit 1 of chapter 10

A tilted set of data

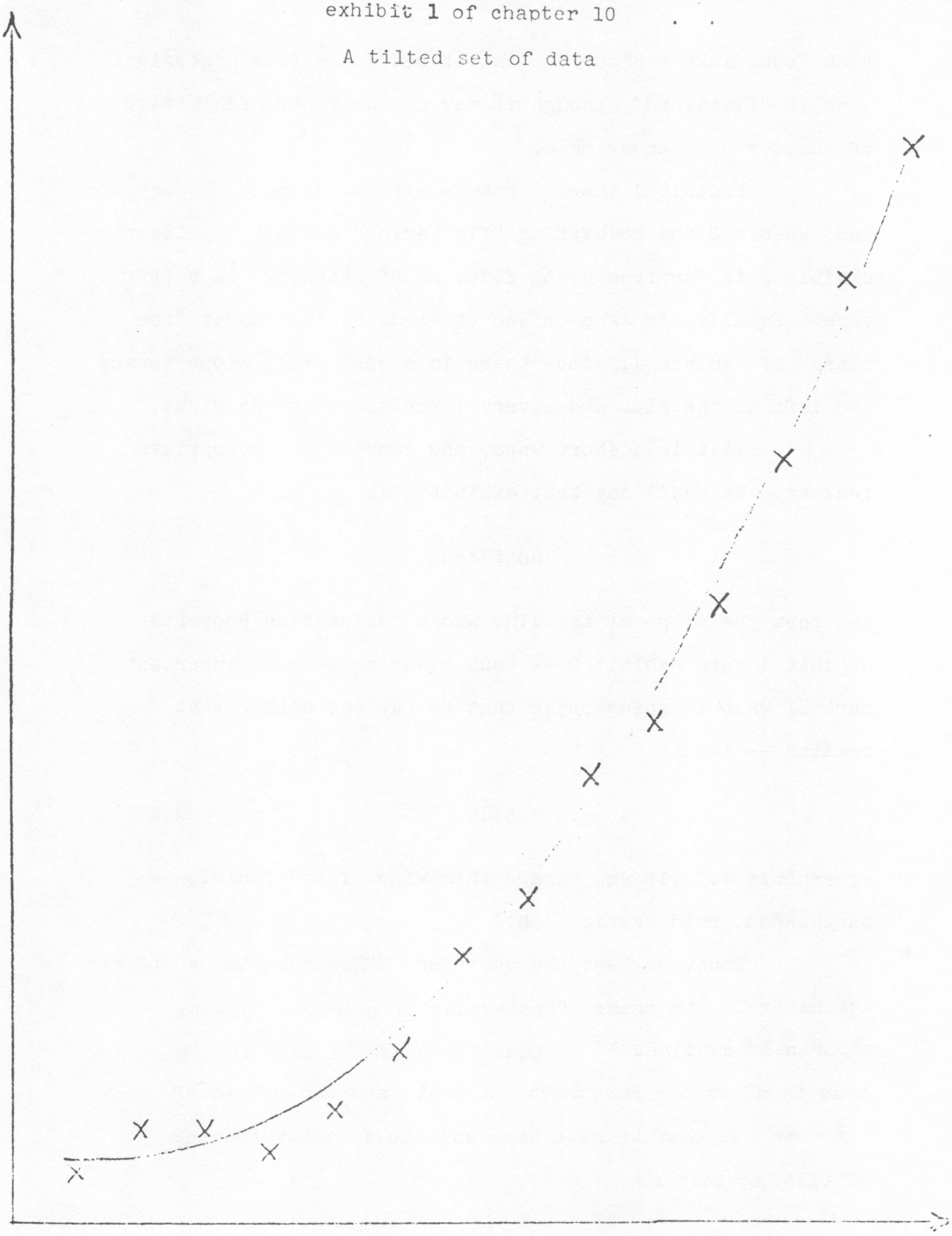
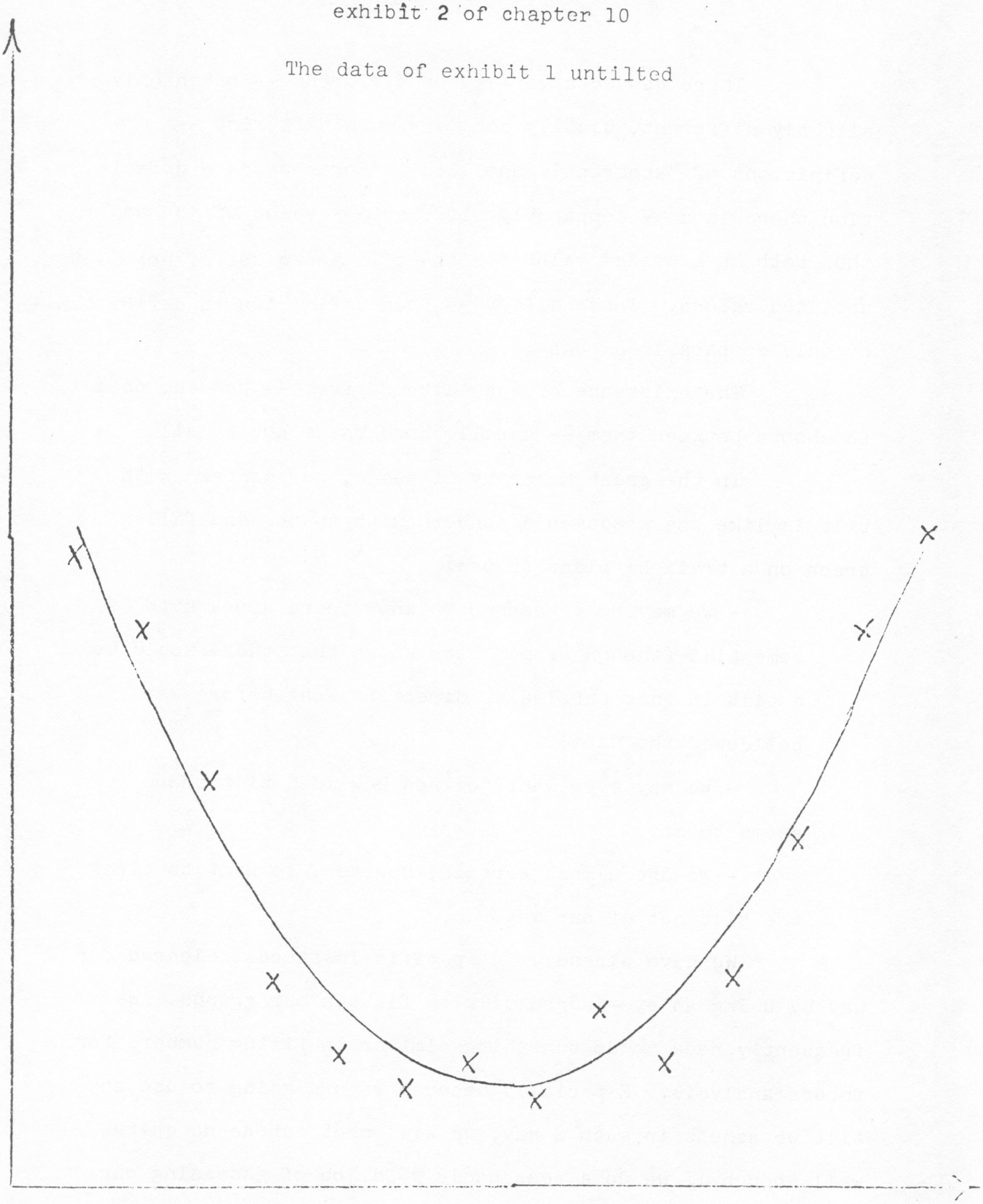


exhibit 2 of chapter 10

The data of exhibit 1 untilted



There can be a variety of different -- often only slightly different, usually not very much different -- definitions of "apparently untilted". For each such definition there is a -- comparably different -- value of b and thus both an assessed value for the tilt and a set of apparently untilted values. These differ -- from definition to definition -- by only comparable amounts.

The existence of such alternatives -- and the need to choose between them -- usually bothers us not at all.

In the great majority of cases, our concern with tilt is like the woodsman's concern with bushes and fallen trees on a trail he plans to use:

- we may be concerned to know there appears to be some tilt (though we may have known that there would be a tilt in this particular direction long before we collected the data).

- we may even want to know how much tilt there seems to be.

- we are almost certain, however, to want to clear the tilt out of our way.

We have already, in specific instances, cleared our way by using an eye-judged tilt to flatten our graphs. We frequently need to do something similar in fixing numbers for future analysis. Especially because we are going to use any tilt we assess in such a way, we will meet our needs quite well enough if we do a reasonably good job of assessing our tilt. We need not worry as to in what way, if any, our assessment might be best possible -- or whether anyone can define "best

* a line for three points *

Suppose we have three (x,y)-points, say (-4,3468), (0,3700), and (4,4138), and we wish to fit a line to them. This means fitting both a level and a tilt. Once we have fitted a tilt, it will be easy to fit a level, since adjustment for level shifts all responses by the same amount.

To fit a tilt, the simplest reasonable thing is to use the two end points of the three. In our case this gives

$$\text{tilt} = \frac{4138-3468}{4-(-4)} = \frac{670}{8} = 83.8 \ .$$

If we now look at the values of (x,y-83.8x), we have

$$(-4,3803) \ , \ (0,3700) \ , \ (4,3803)$$

where $-4\times 83.8 = 335$, $3468+335 = 3803$, $4\times 83.8 = 335$, $4138-335 = 3803$, and, as must be if our arithmetic was right, the end two points have the same values of y-83.8x.

The next step is to fit a level. The mean of these three values of y-83.8x is

$$\frac{3803 + 3700 + 3803}{3} = 3769 \ .$$

For the values of

$$(y-83.8x)-3769 = y(83.8x+3769)$$

we now have +34, -69, and +34. Thus the (x,residual) points

are (-4,34), (0,-69), (4,34), which seem to show no need for further fitting, at least so far as a fitting a line is concerned.

As a second example -- one that is not so symmetrical, -- consider the three points (-22,23.1), (33.8,84.5), (69.5,101.5). Here the end two give a tilt of

$$\frac{101.5 - 23.1}{69.5 - (-22)} = \frac{78.4}{91.5} = .86 \quad .$$

The tilt-adjusted values -- those of y-.86x -- are 42.0, 55.4, 41.3 corresponding to (x,tilt-adjusted y)'s of

(-22,42.0), (33.8,55.4) and (69.5,41.9) .

Again 42.0 and 41.9 agree to within rounding effects.

If again we use the mean, namely

$$\frac{1}{3}(42.0 + 55.4 + 41.9) = 46.4$$

we find, for the residual

$$\text{residual} = (y - .88x) - 46.4 = y - (.86x + 46.4)$$

the values -4.4, 9.0, -4.5 and for the three residual points

(-22,-4.4) , (33.8,9.0) , (69.5,-4.5)

which seem quite satisfactory (except for a tendency to bend).

10A. A simple approach to tilt assessment

* use of three portions *

To assess a tilt most easily, we find the slope of a line through two points. If we start from a plot, we can usually put our finger on two reasonably representative points -- which may or may not be observed points -- with little difficulty. These points are usually -- in a reasonable sense -- "summary points", since they tend to summarize the behavior of the nearby data points.

When we work numerically things are not quite as easy. The simplest thing to do, when working arithmetically, seems to be to identify two batches of observed points, and then calculate a central summary point for each batch. A line through these two summary points can then determine tilt.

While it might seem natural to form the two batches by dividing all the data into two portions, leaving none out, this is usually not either as easy or as effective as is the use of three portions. After we have divided the data into three portions of comparable size, we usually do well to assess tilt from the points that summarize the two end portions.

To look for tilt, then -- either in original observations or in residuals -- we plan to divide the observations into three portions -- low, mid, and high -- and to summarize each of these portions by a central point.
We shall use $(x_L,y_L),(x_M,y_M),(x_H,y_H)$ to denote these points, however found. Our prime concern will be with the corresponding value of

$$\text{apparent tilt} = \frac{y_H - y_L}{x_H - x_L}$$

and we will expect to use a

trial tilt

close to this apparent tilt to make a first adjustment of our data.

It is time again to state clearly that:

- we have not tried to find a best way to assess a tilt, only a good one. Our purpose is to have something that is reasonably easy to understand, reasonably easy to use, and reasonably effective. If we can combine these three properties, we shall be quite happy.

* rule of thumb *

We can avoid much argument about how to divide up our data into portions if we have a rule of thumb to tell us what to do. It does not matter too much which rule we use, so long as it is fairly reasonable. One simple and effective one has these parts:

- neither end portion shall extend more than halfway from one end value of x toward the other end.

- subject to that constraint, and to treating points with the same x alike, each end portion shall come as close to including one-third of all the points as possible, while not including more than one-third.

* choice of origin *

If the values of x are not already convenient, we will routinely replace x by

(constant')(x - constant)

with the two constants so chosen as to make arithmetic easy. It is always easy to allow for this change once the fit has been made.

* using a trial tilt *

Our aim is to flatten the residuals. Moving all of them bodily up or down together to bring them closer to zero can wait. Accordingly, the basic step is to adjust each observed y by subtracting

$$(\text{trial tilt})\cdot(x)$$

where, as we have said, "trial tilt" is to be taken equal -- or close to -- the apparent tilt, and x has been made convenient to work with, thus simplifying arithmetic. It often pays, as we shall see, to choose "trial tilt" to simplify arithmetic, too.

Now that we have residuals, we can replace

$$(x,y)$$

in our thinking by

$$(x, \text{residual})$$

where

$$\text{residual} = y - (\text{trial tilt})\cdot(x)$$

For these new points -- those involving "x" and "residual" -- we want three portions and a central point for each, as before. Labor is saved if we reuse as much as possible. Surely we can keep the same portions; these were chosen on account

of the x's, which are still the same. Clearly we ought to try to keep x_L, x_M, and x_H. We find that we can. All that is needed new, then, are

$$(\text{residual})_L$$

$$(\text{residual})_M$$

$$(\text{residual})_H$$

* taking a 2nd step? *

Once we have come to this point, we can calculate an

$$\text{apparent 2nd tilt} = \frac{(\text{residual})_H - (\text{residual})_L}{x_H - x_L}$$

and do one of three things:

- relax, since the data is so nearly untilted as to require no further action.
- accept a small apparent 2nd tilt as a correction to be added to the trial tilt, but one so small as not to be worth recalculating residuals.
- pick a "trial 2nd tilt" close to our apparent 2nd tilt and find new residuals just as in the initial step.

Which of these three paths we take depends on our judgment of the size of the apparent 2nd tilt.

If needed, new residuals -- 2nd residuals -- are to be found by subtracting

$$(\text{2nd trial tilt})\cdot(x)$$

from each (first) residual.

* a weather eye for bend *

In making this judgment -- and more particularly in deciding how much attention our data demand for humpiness or sag -- we shall lump these curvings together as "bend" -- we will want to look at the apparent relation of the middle portion to the other two. Once we have progressed to residuals, it is usually enough to be simple, to look at

$$\text{mid} - \frac{\text{high} + \text{low}}{2}$$

which stands for

$$(\text{residual})_M - \frac{(\text{residual})_H + (\text{residual})_L}{2}$$

* finishing off *

Once we have come to residuals that seem satisfactorily untilted, bringing them to reasonable balance around zero is easy. We can merely subtract from every residual the mean of the central values for the three portions, namely

$$\frac{(\text{residual})_L + (\text{residual})_M + (\text{residual})_H}{3}$$

Very occasionally, we might want to go back to the individual residuals to make a more subtle adjustment. This possibility will not concern us here.

To get a clear picture of what we may have failed to do, we look at a plot of individual residuals. It often helps to add to this plot, distinctively, three other points representing the summarized behavior of the three portions. The natural choice for these points, which are to be compared with points of the form

$$(x, \text{residual})$$

is to use -- as summary residuals the three points

$$(x_L, (\text{residual})_L)$$

$$(x_M, (\text{residual})_M)$$

$$(x_H, (\text{residual})_H)$$

This usually works quite well.

Occasionally, we find an (x,y) plot to be more effective than an (x, residual) plot. If we are to go back to (x,y) for the data points, for which

$$y = a + bx + \text{residual}$$

it is both natural and appropriate to use

$$(x_L, a + bx_L + (\text{residual})_L)$$

$$(x_M, a + bx_M + (\text{residual})_M)$$

$$(x_H, a + bx_H + (\text{residual})_H)$$

as the summary points.

10B. Untilting with means

Unless the data is reasonably behaved, using means as central values risks perturbation by strays. If there is much data, and we are doing the arithmetic ourselves, means involve extra effort -- especially when compared with trimeans.

On the other hand, however, in eliminating tilt by fitting a straight line, as in a wide variety of other simple processes, the use of means has one great compensation:

- one step is enough, if no errors are made.

For simplicity, then, let us start our actual untilting by using means to find our central points. As we do this, we shall have, for example:

x_L = mean of all observed x's in the low-x portion.

y_H = mean of all observed y's in the high-x portion.

$(\text{residual})_M$ = mean of all residuals in the mid-x portion.

* speed and cost of multiplication *

Exhibit 3 tackles a very small example, comparing speed of multiplication and cost of multiplication as carried out by hand and machine.

With so few observations, we can arrange the calculations somewhat more simply than usually. Panel A presents the data, both in raw form and in logs. (Since the largest/smallest ratios are well over 1,000,000, we are almost sure that logs will be needed.)

exhibit 3 of chapter 10

Speed Compared to Cost for Multiplication by Hand and Machine

A) DATA - raw and logarithmic

Means	**Time*** (microseconds)	**Cost****	**x**	**y**
Pencil and paper	1 minute = 60,000,000	12,500,000	7.78	7.10
Desk calculator	10 seconds = 10,000,000	2,150,000	7.00	6.33
Harvard Mark I	1 second = 1,000,000	850,000	6.00	5.93
ENIAC	10 millisec = 10,000	12,800	4.00	4.11
UNIVAC I	2 millisec = 2,000	4,300	3.30	3.63
UNIVAC 1103	500	1,420	2.70	3.15
IBM 7094	25	132	1.40	2.12
STRETCH	2.5	29	0.40	1.46

* Time to do one multiplication.

** Cost (in dollars) of 125 million multiplications.

x = log of time in microseconds.

y = log of cost in dollars per 125 million multiplications.

B) CALCULATIONS OF PORTIONS -- x-values

	lo	**mid**	**hi**
Halfway	0.40 to 4.08		4.10 to 7.78
Third-count	0.40 to 1.40		7.00 to 7.78
Results	0.40 to 1.40	2.70 to 6.00	7.00 to 7.78

(Cont'd)

exhibit 3 (cont'd.)

C) COMPUTATION ACROSS PORTIONS -- bold entries from panel D

		for portions; basis means			*	apparent	trial
	At x =	0.90	4.00	7.39	(6.49)		
Adjusted:						(tilts)	
	raw y	1.79	4.205	6.715	4.925	.76	.76
	1st adjusted	1.106	1.165	1.098	-.008	-.001	polish
					3	(levels)	
	(same)				[3.369]	1.123	1.120
	residual	-.014	.045	-.022		.003	polish
Fitted:							
.76x	tilt	.684	3.040	5.616			
+ 1.120	level	1.120	1.120	1.120			
.76x + 1.120	total	1.804	4.160	6.736			
	(residual)	(-.016)	(.045)	(-.022)			
same + res'l	restored y	1.790	4.205	6.714			
Polished: (optional)							
(.760x + 1.120)	(total fit)	(1.804)	(4.160)	(6.736)			
-.001x	tilt polish	-.001	-.004	-.007			
+ .004	level for tilt	.004	.004	.004			
+ .003	level polish	.003	.003	.003			
.759x + 1.127	polished fit	1.810	4.163	6.736			
	polished res'l	-.020	.042	.022			

Translated:

y = log of cost ($ per 125 million multiplications) has been fitted by .759(log of multiply time in microseconds) + 1.127

(*) Left column: Value for hi-x portion minus value for low-x portion
Right column: Sum of values for three portions

exhibit 3 (cont'd.)

D) CALCULATION OF ADJUSTED VALUES -- .76 and 1.12 from Panel C

			tilt		level	
	x	y	.76x	Diff	1.12	Diff
	7.78	7.10	5.913	1.197	1.120	.067
	7.00	6.33	5.320	1.010	1.120	-.110
	6.00	5.93	4.560	1.370	1.120	.250
	4.00	4.11	3.040	1.070	1.120	-.050
	3.30	3.63	2.508	1.122	1.120	.002
	2.70	3.15	2.052	1.098	1.120	-.022
	1.40	2.12	1.064	1.056	1.120	-.064
	0.40	1.46	.304	1.156	1.120	.036
means	7.39	6.715		1.098		-.022
by	4.00	4.205		1.165		.045
portion	0.90	1.79		1.106		-.014

E) EXPLANATION of STANDARD CALCULATION ACROSS PORTIONS

All entries in "Adjusted" section come from calculations for separate portions.

- There will be as many "1st adjusted", "2nd adjusted", etc. as in the portion calculations.

- "polish" in the "(trial)" column indicates no further trial tilt or trial level referred to the portion calculations.

- "same" indicates that now we look at the same 3 values for level rather than for tilt.

All early entries in "Fitted" section come from trial tilts or trial levels, as shown by "formula" entries.

- there will be as many tilt adjustments as have been made in the portion calculations (two or more can be combined for the mid-x portion).

exhibit 3 (cont'd.)

- "total" is the sum of the entries directly above.

- "restored y" is the sum of "total" and "(residual)".

Entries in "polished" have mixed origins.

- "(total fit)" is repeated.

- "tilt polish" comes from still unsatisfied tilt in "Adjusted".

- "level for tilt" makes the mean value of the sum of "tilt polish" and "level for tilt" equal zero.

- "level polish" comes from still unsatisfied level in "Adjusted".

- "polished fit" is sum of entries directly above.

- "polished residual" comes from "residual" by SUBTRACTING the sum of the correction terms.

- the identity

$$\text{total fit} + \text{residual} \equiv \text{restored } y$$

is preserved in the new form

$$\text{polished fit} + \text{polished residual} \equiv \text{restored } y \ .$$

The "(*)", "apparent", and "(trial)" columns are used for multiple purposes, as shown in footnote.

(cont'd.)

exhibit 3 (cont'd.)

F. STEP-by-STEP CALCULATION ACROSS PORTIONS -- entries under "Adjusted" from panel G

	for portions, basis means			(*)	apparent	trial
At x =	0.90	4.00	7.39	(6.49)	(see *)	
Adjusted:						
Raw y	1.79	4.205	6.715	(4.925)	.76	(.75)
1st adjusted	1.115	1.205	1.173	(.058)	.009	(.01)
2nd adjusted	1.106	1.165	1.098	(-.008)	-.001	polish
same				[3.369]	1.123	(1.12)
residual	-.014	.045	-.022			
Fitted:						
1st tilt	.675	3.000	5.542		.75x	
2nd tilt	.009	.040	.074		.01x	
level	1.12	1.12	1.12			1.12
total	1.804	4.160	6.736		.75x +	1.12
(residual)	(-.014)	(.045)	(-.022)			
restored y	1.790	4.205	6.714			

Polished:
Left to reader

Translated:
Left to reader

(*) In ( ), value for hi-x portion minus value for lo-x portion -- followed by tilts, divisor 6.49.

In [ ], sum of values for 3 portions -- followed by levels, divisor 3.

(cont'd.)

10-15b

exhibit 3 (Cont'd)

G) STEP-BY-STEP CALCULATION OF ADJUSTED VALUES -- bold numbers from panel F

			1st tilt		2nd tilt		level	
	x	**y**	**.75x**	**Diff**	**.01x**	**Diff**	**1.12**	**Diff**
	7.78	7.10	5.835	1.265	.078	1.187	1.12	-.067
	7.00	6.33	5.25	1.080	.070	1.010	1.12	-.110
	6.00	5.93	4.50	1.43	.060	1.370	1.12	+.250
	4.00	4.11	3.00	1.11	.040	1.070	1.12	-.050
	3.30	3.63	2.475	1.155	.033	1.122	1.12	+.002
	2.70	3.15	2.025	1.125	.027	1.098	1.12	-.022
	1.40	2.12	1.05	1.07	.014	1.056	1.12	-.064
	0.40	1.46	.30	1.16	.004	1.156	1.12	.036
Means	7.39	6.715		1.173		1.098		-.022
by	4.00	4.205		1.205		1.165		.045
portion	0.90	1.79		1.115		1.106		-.014

H) SOURCE

1966. Digital Computer Needs in universities and colleges. Publication 1233 National Academy of Sciences - National Research Council. Table 25 on page 105.

Panel B applies the rule of thumb to determine the three portions.

The remainder of the calculation -- which, as a matter of good practice, we carry to one more decimal than the data -- falls into two sections, each of whose results are repeatedly needed in the other part:

- calculations across portions, here in panel C.
- calculations for individual values, here in panel D. When we have more data, this section will be broken up into separate panels, one for each portion.

(We will come to panels E, F, and G shortly.)

The first steps are in panel D, where the means of x and y are found for each portion. The next steps are in panel C, where the difference in x means from hi to lo is taken as the denominator for calculating apparent tilts, and where the corresponding difference in y means leads to the first apparent tilt. In panel C the trial tilt is taken as 0.76. The next step is to return to panel D, adjust each value by subtracting .76x, and find the means by portions of these adjusted values.

We now return to panel C, bringing 1.098, 1.165, and 1.106 with us. We recalculate an apparent tilt, partly as a check, finding -.001, which is much too small for us to wish to use it in improving adjusted values and residuals. (We can -- if we wish -- use it later to polish the fit.) We can now sum the three values we brought over, and divide by 3, finding an

apparent level of 1.123. For convenience we try adjusting for a level of 1.12.

This adjustment is made in panel D, from which we now return with -.014, .045, and -.022. The apparent tilt is the same as before, and the apparent level is what we might have expected, namely

$$1.123 - 1.12 = .003 \quad .$$

Together, these agreements check some of our arithmetic.

Our adjusted values in panel D -- the last "Diff" column -- are clearly good enough residuals. We could polish them, but to no real good.

Our fit is probably good enough too, but since we can polish it even more easily, we shall do so. The row marked "tilt polish" gives the values of -.001x, the remaining apparent tilt, at x = 0.90, 4.00 and 7.39. The average of these tilt polishes, here -.004, has to be compensated for, as does the +.003 found above. The rows marked "level polish" do this. Adding all three polishes to the total fit gives the polished fit. Subtracting them from the residuals gives the polished residuals.

The right-hand columns show the formulas for the various tilts, levels and combinations.

If we want to plot summary residuals, we would use

(0.90, -0.020), (4.00, +0.042), (7.39, -0.022)

while if we wanted to plot summaries of the fit, we would use

(0.90, 1.810), (4.00, 4.163), (7.39, 6.736) .

The results of adding together these two kinds of summaries

(0.90, 1.790), (4.00, 4.205), (7.39, 6.714)

agree exactly with the means by portion. This is only because we have been untilting with means. When we use more resistant methods, we will not find such precise agreement.

Panel E gives a more general account of the standard form for this calculation.

* step-by-step untilting *

When we come to using resistant methods, we will find that one step of tilt removal is often not enough -- not all the middle adjusted values after the first untilting need come from middle y's, for example. We might as well get used to untilting in stages now, particularly since it need not be more work.

Panels F and G repeat panels C and D with one exception: the 1st trial tilt is .75 -- not .76 -- so we need a 2nd trial tilt of .01. The final answer is the same, as it should be.

Why did we say that it need be no more effort? Let us look at the detailed arithmetic. For 3/4 + .01 it runs

$$\frac{3}{4}(4.00) = 3.00$$

$$4.11-3.00 = 1.11$$

$$0.01(4.00) = .04$$

$$1.11-.04 = 1.07$$

For .76 it would run, breaking down the multiplication

$$.7(4.00) = 2.80$$

$$.06(4.00) = .24$$

$$2.80+.24 = 3.04$$

$$4.11-3.04 = 1.07$$

Some might find this argument stronger if we had subtracted first $.7x$ and then $.06x$ or first $.8x$ and then $-.04x$. However we argue, the conclusion is the same.

Taking "extra" steps by subtracting "round" multiples of x does not cost appreciable extra effort.

* a snare *

Notice now, if you have not noticed before, that -- since we used means -- we could have carried out the calculations of panel C using only x and y means from panel D.

Specifically, we could have used

raw y	1.79	4.205	6.715
fitted tilt	.684	3.040	5.616
difference	1.106	1.165	1.099
1st adjusted	1.106	1.165	1.098
fitted level	1.120	1.120	1.120
difference	-.014	.045	-.022 .

Differences in rounding would have given 1.099 in place of 1.098 -- all else would be the same.

This would NOT have been a good thing, because had we taken this easy way we would not have calculated individual residuals from our fitted line. Sometimes almost all the value of the fit is in these residuals.

Moreover, this only happens when we fit by means. If we use a more resistant way of fitting, this shortcut -- fortunately -- does not work.

Such shortcuts have, unfortunately, seduced many analysts away from the more useful results to be found using resistant tools. We need to avoid them.

* looking at the residuals *

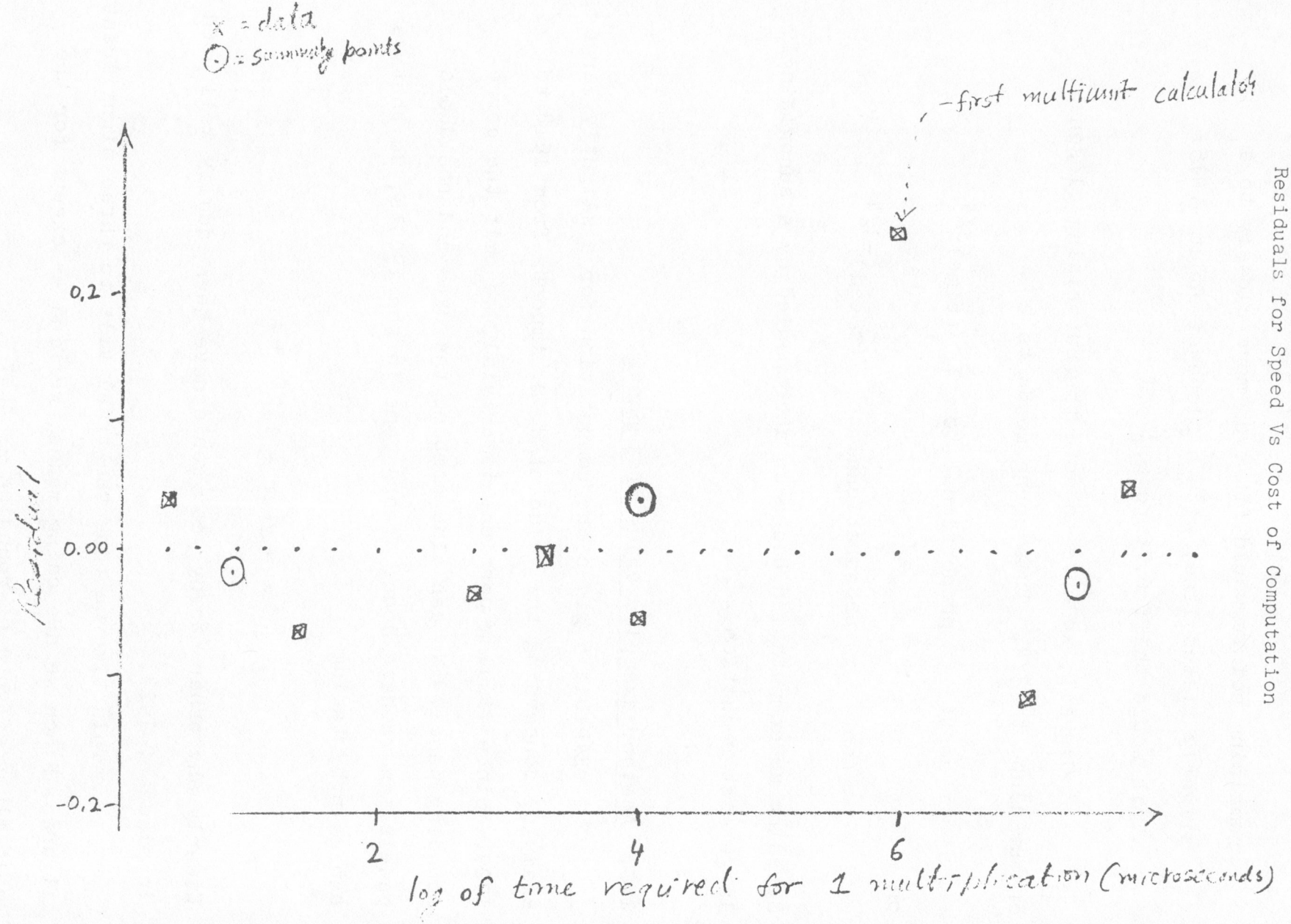
Exhibit 4 shows the residuals and the summary residuals for the three portions. We see:

- one residual that appears quite out of line with the others. (That the first multi-unit calculator should show a larger cost than would be interpolated from the other points is far from surprising.)

- an apparent bend in which the point for the middle portion is above the line joining the points for the other portions. (In evaluating this appearance, we must take due notice of the fact that the one outstanding residual contributes to the means which specify where this summary point falls. Clearly there is a strong case for an analysis in which we use resistant summaries.)

- the right-most point, which is for pencil-and-paper calculation, also seems a little high. If both it and the obviously high point were set aside, the

exhibit 4 of chapter 10
Residuals for Speed Vs Cost of Computation



remaining points would fit even more closely to a slightly tilted line, with a slope of about -.015. (This would shift 0.76 to 0.745.)

Overall, we have to note the surprising closeness of our slope to 3/4. This corresponds to

$$\text{cost} \quad \text{proportional to} \quad (\text{time})^{3/4}$$

or

$$\text{cost} \quad \text{proporational to} \quad (\text{speed})^{-3/4}$$

Speeding up computation has been accompanied by a surprisingly large cost reduction.

10C. <u>Another example, value of imports</u>

Exhibit 5 gives the computation for a slightly larger example, concerning the value of U. S. imports from 1918 to 1929. Since finding portions is trivial, we omit the corresponding panel. We can also combine the data and residuals panels, as we have done. Years come in the 1900's, so that we can gain by taking

$$x = \text{year} - 1924$$

Clearly any other nearby year could have been equally well used in place of 1924.

This time we took the trial tilt to three significant figures. Since we are using means, we find -- except for the most minor effects of rounding -- that

exhibit 5 of chapter 10

Fitting a Line to U. S. Imports, 1918-1929
(y = value in millions of dollars)

A) DATA, ADJUSTED VALUES, RESIDUALS -- see panel B for intervening calculations

year		value	1st tilt		level	
raw	x	y	63.2x	Diff	3882	Diff*
1918	-6	3031	-379.2	3410.2	3882	-472
1919	-5	3904	-316.0	4220.0	3882	+338
1920	-4	5278	-252.8	5530.8	3882	+1649
1921	-3	2509	-189.6	2698.6	3882	-1183
1922	-2	3113	-126.4	3239.4	3882	-643
1923	-1	3792	-63.2	3855.2	3882	-27
1924	0	3609	0	3609.0	3882	-273
1925	1	4227	63.2	4163.8	3882	+282
1926	2	4431	126.4	4304.6	3882	+423
1927	3	4185	189.6	3995.4	3882	+113
1928	4	4091	252.8	3838.2	3882	-44
1929	5	4039	316.0	3723.0	3882	-159

* Residuals. As the final step, obviously no need to carry extra decimal in view of the sizes of the residuals.

(Cont'd)

exhibit 5 (cont'd.)

B) CALCULATION ACROSS PORTIONS

	for portions; basis means			(*)	apparent	(trial)
At x =	**-4.5**	**-0.5**	**3.5**	(8)	(see *)	
Adjusted:						
raw y	**3680.5**	3685.2	4186.5	(506)	63.2	(63.2)
tilt adjusted	3964.9	3716.8	3965.3	(0.4)	.05	polish
same				[11647.0]	3882.3	(3882)
residual	82.9	-165.2	83.3	[1.0]	.3	polish
					formula	
Fitted:						
tilt	-284.4	-31.6	221.2		63.2x	
level	3882.	3882.	3882.		+ 3882	
total	3597.6	3850.4	2103.2		63.2x + 3882	
(residual)	(82.9)	(-165.2)	(83.3)			
restored y	3680.5	3685.2	4186.5			

Polished:
Left to reader

Translated:
value of imports (in million $) ∿63.2 (year-1924) + 3882

Problem: Calculate the polished fit (see panel C of exhibit 3)

C) SOURCE

10-25

$$(\text{1st adjusted})_L = (\text{1st adjusted})_H$$

the actual values being 3964.9 and 3965.3. Thus the 1st adjustment has been fully effective; there is no need for a 2nd step.

There is clearly little need for a polishing, so we leave this to the reader. There is, however, need to translate the x-values for the summary points into years. This is done at the foot of panel B.

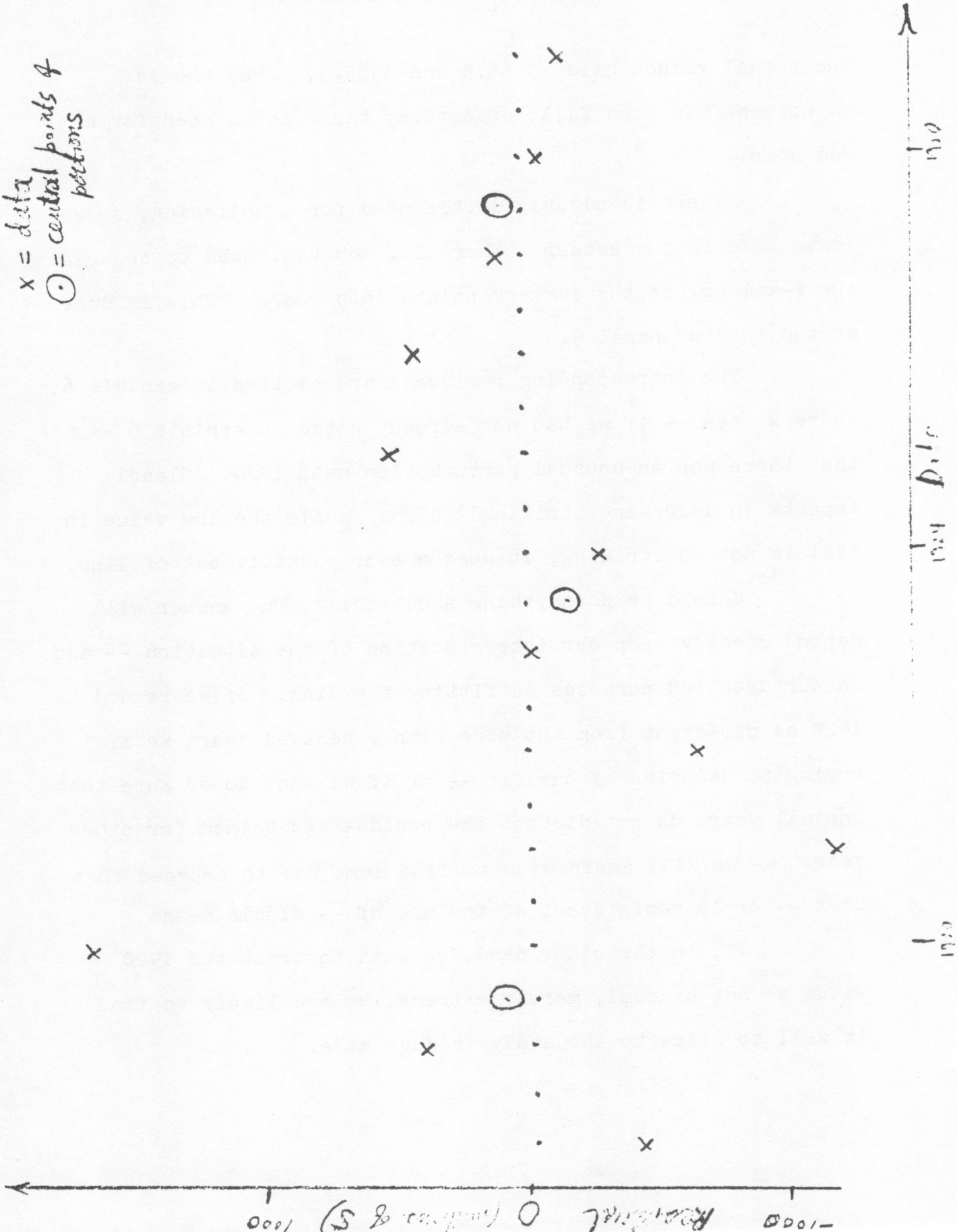
The corresponding residuals are plotted in exhibit 6, where we see -- if we had not already noted in exhibit 5 -- that there was an unusual perturbation near 1920. Clearly imports in 1920 were strikingly high. While the low value in 1921 is not so striking, it does appear possibly out of line.

Should we do anything about this? The answer will depend greatly upon our interpretation of the situation -- and on our detailed purposes in fitting the line. If we regard 1920 as different from the more simply behaved years we are trying to describe by our fit -- or if we want to be sure that unusual years do not distort the residuals obtained for other years -- we will again wish to find some way to proceed that uses -- or is reminiscent of the use of -- middle means.

If, on the other hand, we want to treat the 1920 value as not unusual, merely extreme, we are likely to find it well to stick to the analysis just made.

exhibit 6 of chapter 10

The Residuals From Line Fitted in exhibit 5



10D. That middle-mean flavor

To use middle means instead of means to find the coordinates of our summary points is easy -- and workable. We can think of complicated ways to find summary points, but since we plan to take a 2nd step whenever necessary, we can safely avoid complication and proceed simply.

What we are trying to reach is

$$(y-bx)_H = (y-bx)_L$$

where these summary values are now defined by middle means. However we get there, all will then be well.

Let us, then, begin by using midmeans on the data of the previous examples. Exhibit 7 gives the modified results for the computing example. As noted in panel A, end-portion summaries and tilts are unaffected because there are so few points.

Notice that "restored y" for the middle portion does NOT reproduce the "raw y" value exactly. We expect such differences wherever middle means are being effective.

Note also that

$$\text{mid} - \frac{1}{2}(\text{high} + \text{low})$$

has been shrunk from

$$.063 = .042 - \frac{1}{2}(-.020-.022)$$

exhibit 7 of chapter 10

Changes in the analysis of exhibit 3 due to the use of middle means

A) IMPORTANT NOTE

Because only 8 (x,y) points are involved, thereaare only two points in each end portion. Thus changing to middle means leaves the summary values for end portions the same. Accordingly tilts are unchanged. With four values in the middle portion, we clearly use medians there. (Note that the median x for the midposition is 3.65, whereas the mean was 4.00.)

B) What panel E of exhibit 5 becomes -- some changed values bold

	for portions; basis means			*	apparent	trial
At x =	**0.90**	**3.65**	**7.39**	(6.49)	(see *)	
Adjusted:						
raw y	1.79	**3.87**	6.715	(4.925)	.76	(.75)
1st adjusted	**1.115**	**1.140**	**1.173**	(.058)	.009	(.01)
2nd adjusted	1.106	**1.110**	1.098	(-.008)	-.001	polish
same				[3.314]	1.105	(1.10)
residual	.006	.000	-.002	[.004]	.001	polish
					formulas	
Fitted:						
1st tilt	.675	2.738	5.542		.75x	
2nd tilt	.009	.036	.074		.01x	
level	1.100	1.100	1.100		+ 1.100	
total	1.784	3.874	6.716		.76x + 1.100	
(residual)	(.006)	(.000)	(-.002)			
restored y	1.790	3.884	6.714			
Polished:						
(total fit)	1.784	3.874	6.716		.760x + 1.100	
tilt polish	-.001	-.004	-.007		-.001x	
level for tilt	+.004	.004	.004		+ .004	
level polish	+.001	.001	.001		+ .001	
polished fit	1.790	3.875	6.721		.759x + 1.105	
polished residual	.002	-.001	.000			
	*****	*****	*****			

10-29

exhibit 7 (cont'd.)

(*) In ( ), value for hi-x portion minus value for lo-x portion -- followed by tilts, divisor 6.49.

In [ ], sum of values for 3 portions -- followed by levels, divisor 3.

Examples:

1.140 is the median of 1.43, 1.11, 1.155 and 1.125 (see panel F of exhibit 3);
1.110 is the median of 1.370, 1.070, 1.122 and 1.098 (see same panel).

C) INDIVIDUAL RESIDUALS

x	residuals from	
	76x + 1.10	.759x + 1.105
7.78	.087	.090
7.00	-.090	-.088
6.00	.270	.271
4.00	-.030	-.031
3.30	.022	.020
2.70	-.002	-.064
1.40	-.044	-.048
0.40	.056	.051

10-29a

to

$$.00 = -.001 - \frac{1}{2}(.002+.000)$$

Clearly the midmean analysis shows no appreciable trace of bend. Exhibit 8 shows the individual and summary residuals, and makes it clear that all relevant trace of bend has disappeared from the summary residuals.

Let us turn next to the analysis of the imports example by midmeans. Exhibit 9 shows the calculations, with various detailed choices altered. Here we took x = year - 1923.5. This gives more symmetric values than our previous choice, but the values of x are now half-integers rather than integers. We have also chosen to fit our tilt one digit at a time, as forecast earlier.

In finer detail, note that we have chosen to show portion midmeans in panel A, and to use the w-technique to avoid rewriting small batches when finding middle means. Use of this technique is wholly optional. It does, however, save effort as explained in exhibit 10.

The final fit is now

$$74.5(\text{year}-1924) + 3830$$

to compare with the fit by means of

$$63.2(\text{year}-1924) + 3882$$

Allowing the high value in 1920 to have a weaker effect has both increased the tilt and lowered the level, as we should have expected.

exhibit 8 of chapter 10

Residuals From the Midmean Analysis
(compare exhibit 4)

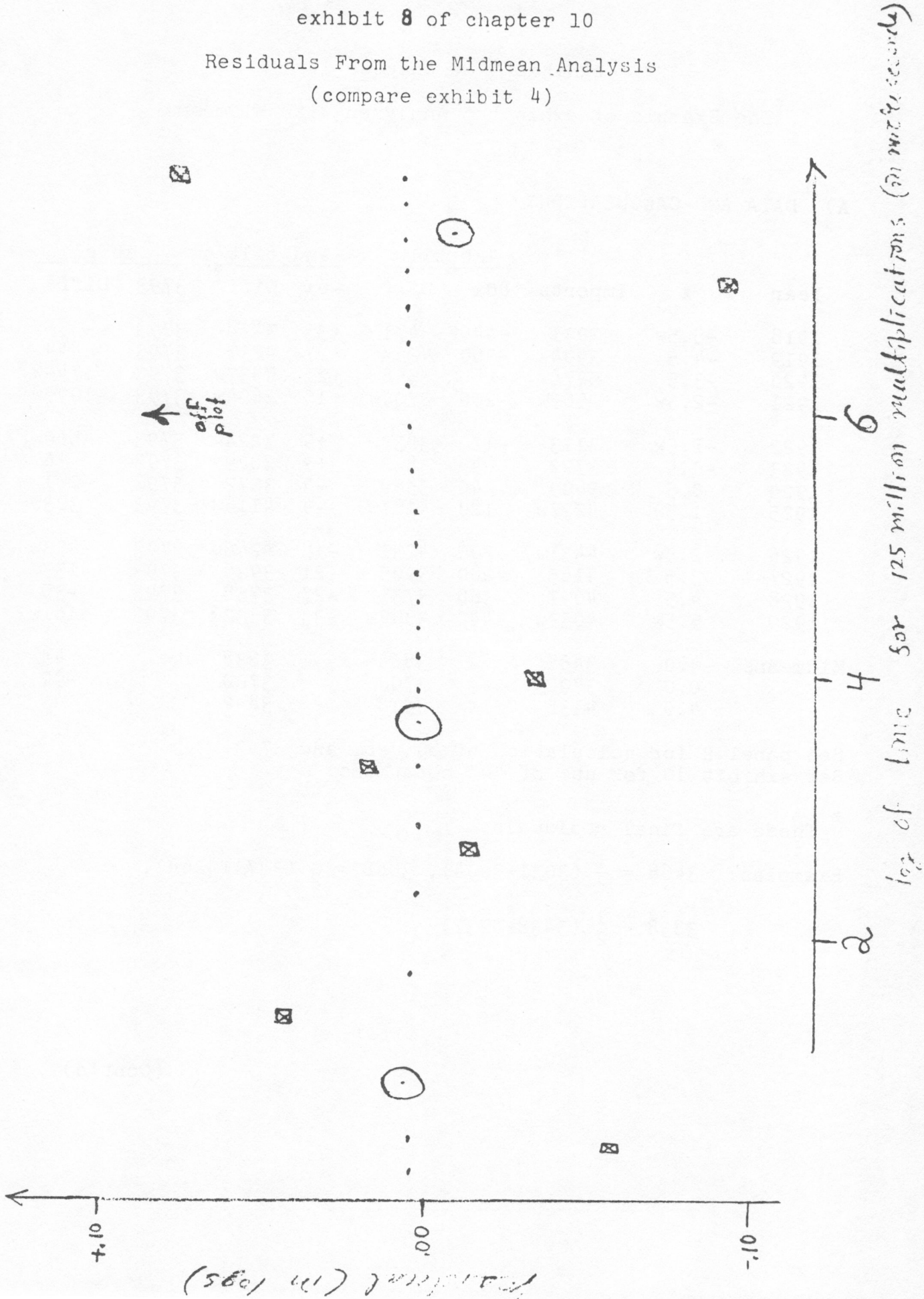


exhibit 9 of chapter 10

The Example of exhibit 5 Analyzed With Midmeans

A) DATA AND CALCULATIONS

			1st tilt		2nd tilt		level	
Year	**x**	**Imports**	**80x**	**Diff**	**-6x**	**Diff**	**3793**	**Diff***
1918	-5.5w	3031	-440	3471	+33	3438	3793	-355
1919	-4.5	3904	-360	4264	+27	4237	3793	444
1920	-3.5	5278w	-280	5558w	+21	5537w	3793	1744w
1921	-2.5w	2509w	-200	2709w	+15	2694w	3793	-1099w
1922	-1.5w	3113w	-120	3233w	+9	3224w	3793	-569w
1923	-0.5	3792	-40	3832	+3	3829	3793	36
1924	0.5	3609	40	3569	-3	3572	3793	-221
1925	1.5w	4227w	120	4107w	-9	4116w	3793	323w
1926	2.5w	4431w	200	4231w	-15	4246w	3793	453w
1927	3.5	4185	280	3905	-21	3926	3793	133
1928	4.5	4091	360	3731	-27	3758	3793	-35
1929	5.5w	4039w	440	3599w	-33	3632w	3793	-161w
Midmeans	-4.0	3468		3868		3838		44
	0.0	3700		3700		3700		-92
	4.0	4138		3818		3842		49

See panel B for calculation of 80, -6, and 3793.
See exhibit 10 for use of "w" notations.

* These are final residuals.

Examples: $3468 = \frac{1}{2}(3031+3904)$, $3868 = \frac{1}{2}(3471+4264)$,

$3838 = \frac{1}{2}(3438+4237)$

(Cont'd)

exhibit 9 (cont'd.)

B) CALCULATION ACROSS PORTIONS

	for portions; basis means			(*)	apparent	(trial)
At x =	-4	0	4	(8)	(see *)	
Adjusted:						
y-value	3468	3700	4138	(670)	83.8	(80)
1st adjusted	3868	3700	3818	(-50)	-6.2	(-6)
2nd adjusted	3838	3700	3842	(4)	0.5	polish
(same)				[11380]	3793	3793
residual	45	-93	49	[1]	0.3	polish
Fitted:	****	****	****		formulas	
1st tilt	-320	0	320		80x	
2nd tilt	24	0	-24		-6x	
level	3793	3793	3793		+ 3793	
total	3497	3793	4089		74x + 3793	
(residual)	(45)	(-93)	(49)			
restored y	3542	3700	4138			
	****	****	****			
Polished:						
(total fit)	(3497)	(3793)	(4089)		(74.0x + 3793)	
tilt polish	-2	0	2		0.5x	
level polish	0.3	0.3	0.3		+ 0.3	
polished fit	3495.3	3793.3	4091.3		74.5x + 3793.3	
polished resid.	46.7	-93.3	46.7			
	****	****	****			

(cont'd.)

10-33

exhibit 9 (cont'd.)

Translated:

value of imports ∿74.5 (year-1923.5) + 3793.3

or, if we prefer:

value of imports ∿74.5 (year-1926) + 3830.5.

(*) In ( ), value for hi-portion minus value for lo portion -- followed by tilts, divisor 8.

In [ ], sum of values for 3 portions -- followed by levels, divisor 3.

Note: 3868 is quite different from 3468 - (-320) = 3788.

When we look at panel A, whence comes 3868, we can easily see why the change is by - (-400) and not by - (-320).

exhibit **10** of chapter 10

The Use of "w" Marks in Calculating Middle Means of Small Batches Which Are Already Written Down in Fixed Order

A) CONVENTIONAL CALCULATION

Given	**Put in order**	**Pull out middle**	**Form mean**
3031	2509		
3904	3031	3031	3468
5278	3904	3904	
2509	5278		

B) USE OF "w" MARKS

Given	**Find wings**	**Take mean of rest**
3031	3031	3468
3904	3904	
5278	5278w	
2509	2509w	

C) A LARGER EXAMPLE, WHERE "H" FOR HINGES IS ALSO NEEDED

Given	**Find wings**	**and hinges**	**Take mean of rest**
3113	3113w	3113w	
3792	3792	3792	
3609	3609	3609H	
4227	4227	4227H	4024
4431	4431w	4431w	
4185	4185	4185	
4095	4095	4095	

$$4024 = \frac{1}{3}(3792+4095+4185)$$

D) COMMENT

This device works well in small batches. A similar device works well when seeking the trimean of a somewhat larger batch. There we need to find and mark the median, too. To keep matters relatively clear, the following marking is recommended:

w	wing value
H	hinge
h	one of two values whose mean is a hinge
M	median
m	one of two values whose mean is the median

Exhibit 11 shows the individual and summary residuals, of course taken from the unpolished fit 74x+3493. We see that:

- the midmean for the middle portion is somewhat below those for the end portions. In view of the scatter of the individual residuals, this dip seems not worthy of great attention.

Again the use of middle means seems to have cleared up the analysis at no extra charge. We cannot expect this to happen always -- it should not always happen, sometimes strays can and do conceal real bends -- but when it does, we can welcome its appearance. In numerical terms

$$\text{mid} - \frac{1}{2}(\text{high} + \text{low})$$

is altered from

$$-248 = -165 - \frac{1}{2}(83+83) \qquad \text{(exhibit 5)}$$

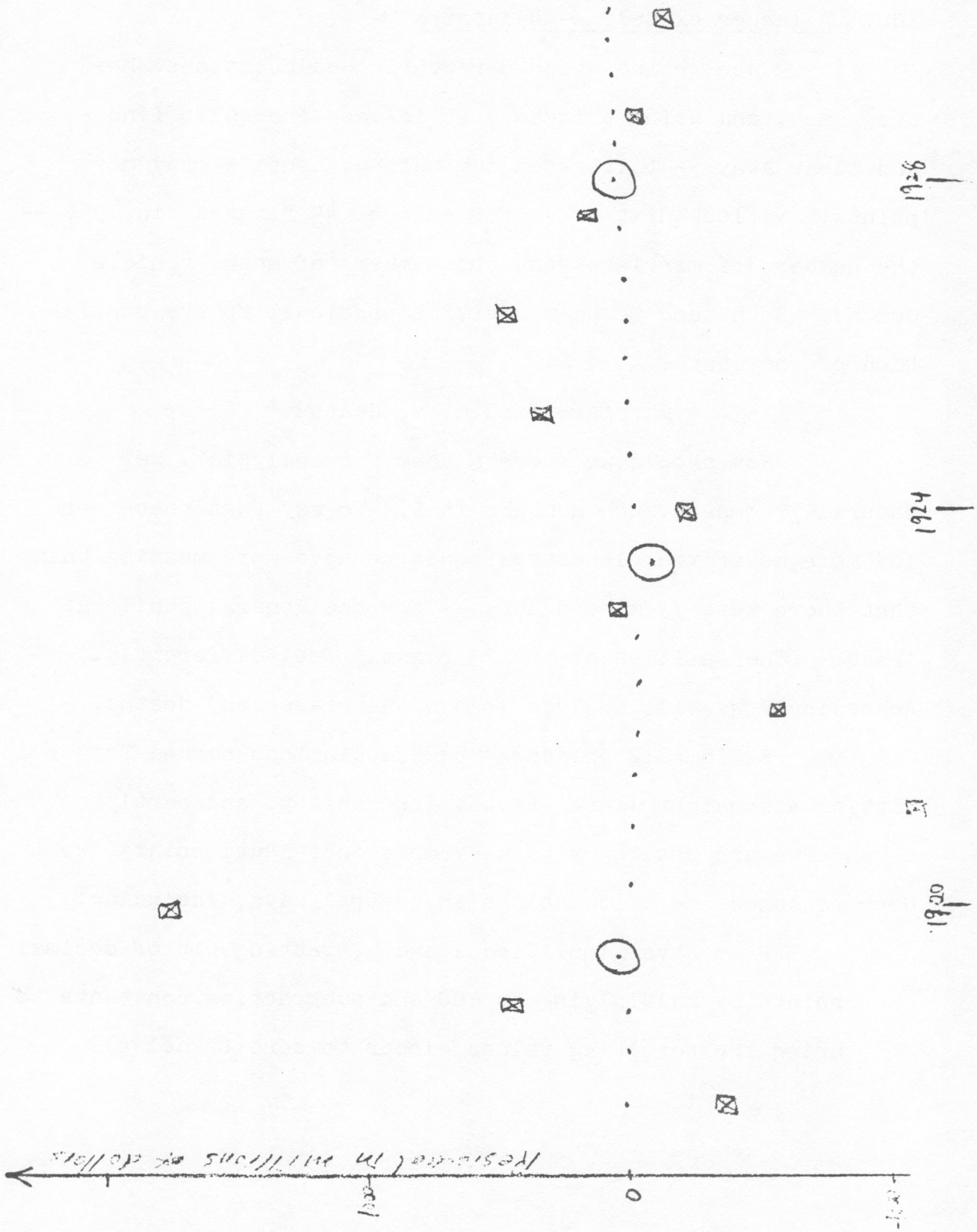
to

$$-139 = -92 - \frac{1}{2}(47+47) \qquad \text{(exhibit 9)}$$

being almost cut in half.

exhibit 11 of chapter 10

Residuals From the Analysis of exhibit 9



10E. A larger example -- 49 states

Once we are used to repeating calculations where necessary, and using trimeans, it is easy enough to find -- and clear away -- tilts when our batches contain many more points. We look next at -- for each of 49 "states" in 1958 -- the numbers of marriages and the numbers of motor vehicle deaths. Both tend to be crudely proportional to the population of the state.

* marriages and m. v. deaths *

How should we express them for analysis? We naturally compare such numbers in %. To say that there were 10% more motor vehicle deaths tends to have more meaning than that there were 73 more deaths -- for the general public at least. (The families of the 73 rightly feel differently.) Accordingly we will analyze logs of marriages and deaths.

Exhibit 12 presents the calculations needed for fitting a straight line. Partly for variety, and partly because we are dealing with many more individual points, we have reshaped the calculations in several ways, including:

- we have simplified x and y, getting rid of decimal points by multiplying by 100 and subtracting constants to bring the resulting values closer to zero (Panel A).

exhibit 12 of chapter 10

Marriages and Motor Vehicle Deaths by State, 1958

A) DATA AND LOGS

	Marriages	Motor vehicle deaths	x	y
Ala	24,444	852	39	93
Ariz	9,917	508	0	71
Ark	15,574	444	19	65
Calif	96,330	3510	98	155
Colo	14,688	396	17	60
Conn	16,879	251	23	40
Del	2,311	84	-63	-8
DC	8,094	62	-9	-21
Fla	35,243	1134	25	105
Ga	45,863	956	66	98
Ida	9,520	270	-2	43
Ill	82,860	1886	92	128
Ind	41,226	1056	62	102
Iowa	25,101	598	40	78
Kans	15,481	554	19	74
Kent	26,095	789	42	90
La	21,447	844	33	93
Maine	7,104	205	-15	31
Maryl	41,403	505	62	70
Mass	45,959	590	66	77
Mich	53,662	1375	73	114
Minn	23,410	707	37	85
Miss	36,198	548	56	74
Mo	33,976	975	53	99
Mont	6,160	193	-22	29
Nebr	10,637	346	3	54
Nev	5,755	190	-24	28
NH	7,078	101	-15	0
NJ	39,113	754	59	88
NM	5,850	408	-23	61
NY	124,513	2118	110	133
NC	27,228	1081	43	103
ND	4,306	155	-37	19
Ohio	65,479	1812	82	126
Okla	33,444	667	54	83
Oreg	9,798	448	-1	65
Penna	64,529	1654	81	122
RI	5,653	74	-25	-13
SC	38,550	610	59	79
SD	5,662	240	-25	38
Tenn	27,947	719	45	86
Tex	89,702	2342	95	137
Utah	6,741	191	-17	28
Vt	3,376	77	-47	-11
Va	36,588	861	56	94
Wash	27,966	573	45	76
WVa	14,213	387	15	59
Wisc	25,073	822	40	91
Wyo	2,945	137	-53	14

x = -400 plus 100 times log of number of marriages, 1958.

y = -200 plus 100 times log of number of motor vehicle deaths, 1958.

exhibit 12 of chapter 10 (Cont'd)

B) STEM-AND-LEAVES -- for all 49 points

for x

-6	3
-5	3
-4	7
-3	7
-2	24355
-1	557
-0	921
0	03
1	9795
2	35
3	937
4	023550
5	639496
6	6226
7	3
8	21
9	825
10	
11	0

for y

-2	1
-1	31
-0	8
0	0
1	94
2	988
3	18
4	03
5	49
6	5015
7	18407496
8	5836
9	3803941
10	523
11	4
12	862
13	37
14	
15	5

C) PORTIONS -- using panel B

halfway: -63 to 23
24 to 110

up to $\frac{1}{3}$ count: -63 to 0
54 to 110

result: -63 to 0
54 to 110

(Cont'd)

exhibit 12 of chapter 10 (Cont'd)

D) LOW-x CALCULATIONS -- x from -63 to 0 -- 16 points

		1st tilt		2nd tilt		level	
x	y	.9x	Diff	-.1x	Diff	46.2	Residual
0	71	0.	71.	0.0	71.	46.2	24.8
-63	-8	-56.7	48.7	6.3	42.4	46.2	-3.8
-9	-21	-8.1	-12.9	0.9	-13.8	46.2	-60.0
-2	43	-1.8	44.8	0.2	44.6	46.2	-1.6
-15	31	-13.5	44.5	1.5	43.0	46.2	-3.2
-22	29	-19.8	48.8	2.2	46.6	46.2	0.4
-24	28	-21.6	49.6	2.4	47.2	46.2	1.0
-15	0	-13.5	13.5	1.5	12.0	46.2	-34.2
-23	61	-20.7	81.7	2.3	79.	46.2	32.2
-37	19	-33.3	52.3	3.7	48.6	46.2	2.4
-1	65	-0.9	65.9	0.1	65.8	46.2	19.6
-25	-13	-22.5	9.5	2.5	7.0	46.2	-39.2
-25	38	-22.5	60.5	2.5	58.0	46.2	11.8
-17	28	-15.3	43.3	1.7	41.6	46.2	-4.6
-47	-11	-42.3	31.3	4.7	26.6	46.2	-19.6
-53	14	-47.7	61.7	5.3	56.4	46.2	10.2

				(cut)		(cut)		(cut)	
-6	3			-1	2	-1	3	-6	0
-5	3	-2	1	-0		-0		-5	
-4	7	-1	31	0	9	0	7	-4	
-3	7	-0	8	1	3	1	2	-3	49
-2	24355	0	0	2		2	6	-2	
-1	557	1	94	3	1	3		-1	9
-0	0921	2	988	4	844893	4	2436781	-0	3134
		3	18	5	2	5	86	0	012
		4	3	6	501	6	5	1	910
		5		7	1	7	19	2	4
		6	15	8	1	8		3	2
		7	1						
				(full)		(full)		(full)	
16		16		16		16		16	
4hH	-31	H	-4	H	37.3	H	34.1	4hH	-12.1
8hM	-22h	M	28	M	48.8	M	45.6	8hM	-0.6
4hH	-12	H	40h	H	61.1	H	57.2	4hH	11.0
tri	-22	tri	23.1	tri	49.0	tri	45.6	tri	-0.6

*

Example: According to the cut stem-and-leaf, the first hinge for the "1st tilt" Diff column falls halfway between "35" and "43". When we restore the decimals, by looking up the column, we find these to be 35.5 and 43.3 so that their mean is 39.4, as entered. The median for the same column lies between "48" and "48", which turn out to be 48.7 and 48.8, and is entered as 48.8. And so on.

(Cont'd)

exhibit **12** of chapter 10 (Cont'd)

E) HIGH-x CALCULATIONS -- x from 54 to 110 -- 16 points

		1st tilt		2nd tilt		level	
x	**y**	**.9x**	**diff**	**-.1x**	**diff**	**46.2**	Residual
98	155	88.2	66.8	-9.8	76.6	46.2	30.4
66	98	59.4	38.6	-6.6	45.2	46.2	-1.0
92	128	82.8	45.2	-9.2	54.4	46.2	8.2
62	102	55.8	46.2	-6.2	52.4	46.2	6.2
62	70	55.8	14.2	-6.2	20.4	46.2	-25.8
66	77	59.4	17.6	-6.6	24.2	46.2	-22.0
73	114	65.7	48.3	-7.3	55.6	46.2	9.4
56	74	50.4	23.6	-5.6	29.2	46.2	-17.0
59	88	53.1	34.9	-5.9	40.8	46.2	-5.4
110	133	99.0	34.0	-11.0	45.0	46.2	-1.2
82	126	73.8	52.2	-8.2	60.4	46.2	14.2
54	83	48.6	34.4	-5.4	39.8	46.2	-6.4
81	122	72.9	49.1	-8.1	57.2	46.2	11.0
59	79	53.1	25.9	-5.9	31.8	46.2	-14.4
95	137	85.5	51.5	-9.5	61.0	46.2	14.8
56	94	50.4	43.6	-5.6	49.2	46.2	3.0

				(cut)		(cut)		(cut)
5	69496	7	0749	1	47	1		-2	52
6	6226	8	83	2	35	2	049	-1	74
7	3	9	84	3	8444	3	91	-0	1516
8	21	10	2	4	56893	4	5059	0	8893
9	825	11	4	5	21	5	4257	1	41
10		12	862	6	6	6	01	2	
11	0	13	37	7		7	6	3	0
		14							
		15	5	(full)		(full)		(full)	
16		16		16		16		16	
4hH	59	H	81	H	30.0	H	35.8	4hH	-10.-
8hM	65	M	100	M	41.1	M	47.2	8hM	1.1
4hH	87	H	127	H	48.7	H	56.4	4hH	10.2
tri	69.5	tri	102	tri	40.2	tri	46.6	tri	0.-

(Cont'd)

exhibit 12 (cont'd.)

F) CALCULATION ACROSS PORTIONS -- lo, mid, and hi

	for portions; basis trimeans			(*)	apparent	trial
At x =	-22	33.8	69.5	(91.5)		
Adjusted:						
raw y	23.1		102.	(78.9)	.862	(.9)
1st adjusted	49.0		40.2	(-8.8)	-.096	(-.1)
2nd adjusted	45.6		46.6	(1.0)	.011	polish
same				⟨92.2⟩	46.1	(46.2)
residual	-0.6	8.0	0.4	[7.8]	2.6	polish
Fitted:					formulas	
1st tilt	-19.8	30.42	62.55		.9x	
2nd tilt	2.2	-3.38	-6.95		-.1x	
level	46.2	46.2	46.2			+ 46.2
total	28.6	73.24	101.80		.8x	+ 46.2
(residual)	(-0.6)	(8.0)	(0.4)			
restored y	28.0	81.24	102.2			
Polished:						
(total fit)	(28.6)	(73.24)	(101.80)		(.800 × 46.2)	
tilt polish	-0.24	0.37	0.76		.011x	
level for tilt	-0.30	-0.30	-0.30			- 0.30
level polish	2.60	2.60	2.60			+ 2.60
polished fit	30.66	75.91	104.86		.811x	+ 48.5
polished residual	-2.66	5.33	-2.66			

Translated:

log (motor vehicle deaths) = .811 log(marriages) - .759

(*) In ( ), value for hi portion minus value for low portion -- followed by tilts, divisor 91.5.

In ⟨ ⟩, sum of values for hi and low portions -- followed by levels, divisor 2.

In [ ], sum of values for all 3 portions -- followed by lev ls, divisor 3.

(cont'd.)

10-42

exhibit 12 (cont'd.)

F) Continued

Notes:

1) See panel A for translation of y and x.

2) 49.0 does NOT equal 42.9 = 23.1 - (-19.8), because 49.0 comes from panel D.

3) Since adjustment of all values by a constant moves the trimean by the same constant, it follows that -0.6 DOES equal 45.6 - 46.2.

G) MID-x CALCULATION -- x from 3 to 53 -- 17 points

		1st and 2nd tilt		level	
x	**y**	**.8x**	**Diff**	**46.2**	**Residual***
39	93	31.2	61.8	46.2	15.6
19	65	15.2	49.8	46.2	3.5
17	60	13.6	46.4	46.2	0.2
23	40	18.4	21.6	46.2	-24.6
25	105	20.0	85.0	46.2	38.8
40	78	32.0	46.0	46.2	-0.2
19	74	15.2	58.8	46.2	12.6
42	90	33.6	56.4	46.2	10.2
33	93	26.4	66.6	46.2	20.4
37	85	29.6	55.4	46.2	9.2
53	99	42.4	56.6	46.2	10.4
3	54	2.4	51.6	46.2	5.4
43	103	34.4	68.6	46.2	22.4
45	86	36.0	50.0	46.2	3.8
45	76	36.0	40.0	46.2	-6.2
15	59	12.0	47.0	46.2	0.8
40	91	32.0	59.0	46.2	12.8

```
0|3
1|9795
2|35
3|937
4|023550
5|3
```

(cut)

```
-2|4
-1|
-0|06
 0|309530
 1|52002
 2|02
 3|8
```

(full)

```
17|
5H|13
9M|37
5H|42
tri|33.8
```

```
17|
5H|  0.8
9M|  9.2
5H| 12.8
tri|  8.0
```

exhibit 13 of chapter 10

Residuals from exhibit 12

$$.011x + 2.6$$

from the fit defining our best individual residuals -- we have plotted the polished fit as well as the residuals. We now see that:

- the apparent bend in summary residuals is not large, in comparison with the scatter of individual residuals, though it may not be negligible.
- changing the middle portion points to "+" instead of "x" causes little or no overall confusion, and is thus probably worthwhile, since we may want to see which points belong to which portion.
- there seems to be a tendency for a few residuals to scatter.

* looking individually *

To look further into all 49 residuals, we turn to exhibit 14, where panel A finds a total of nine residuals "outside". When we turn to the individual outside residuals -- in panel B -- we see:

- D.C. and Maryland low.
- New England states low.
- California and New Mexico high.
- Florida high.

With this much geographic structure, we ought to look at the neighboring states, as in panel C. The neighbors of D.C.

exhibit 14 of chapter 10

Analysis of all 49 residuals of exhibit 12

A) STEM-and-LEAF DISPLAY -- individual residuals cut to integers, as collected from panels D, E and G of exhibit 12.

1	-6	0	0	
.	-5			
.	-4			
3	-3	94	94	
6	-2	542	542	
9	-1	974		974
to 18	-0	655433111		655433111
49 ✓14	0	00011234456899		00011234456899
to 17	1	01111234469		01111234469
6	2	134		134
3	3	029	029	
			(outside)	(inside)

49		dh	15
13H	-4	S	-19
13H	11	S	26

B) The OUTSIDE RESIDUALS

State	Residual	x	y
District of Columbia	-60	-9	-21
Rhode Island	-39	-25	-13
New Hampshire	-34	-15	0
Maryland	-25	62	70
Connecticut	-24	23	40
Massachusetts	-22	66	77
California	30	98	155
New Mexico	32	-23	61
Florida	39	25	105

(cont'd.)

C) GEOGRAPHIC NEIGHBORS -- of the states in B

State	Residual	x	y	Relation
Delaware	-3	-63	-8	neutral
Pennsylvania	11	81	122	opposite
West Virginia	1	15	59	neutral
Virginia	3	56	94	neutral
Maine	-3	-15	31	neutral
Vermont	-19	-47	-11	ALSO LOW *
New York	-1	110	133	neutral
Oregon	19	-1	65	also high *
Nevada	1	-24	28	neutral
Arizona	24	0	71	ALSO HIGH *
Utah	-4	-17	28	neutral
Colorado	0	17	60	neutral
Oklahoma	-6	54	83	neutral
Texas	14	95	137	also high *
Georgia	-0	66	98	neutral
Alabama	16	39	93	also high *

* Five more states that may deserve special attention.

and Md. tend to be low. Maine and New York do not follow the first four New England states, but Vermont seems to. In the West, we pick up Oregon, Arizona, and Texas as also high. In the south we find Alabama as also high.

All this suggests setting aside:

- D.C. and Maryland (low).
- five New England states (low).
- Arizona, California, New Mexico, Oregon, Texas (high).
- Alabama, Florida, (high).

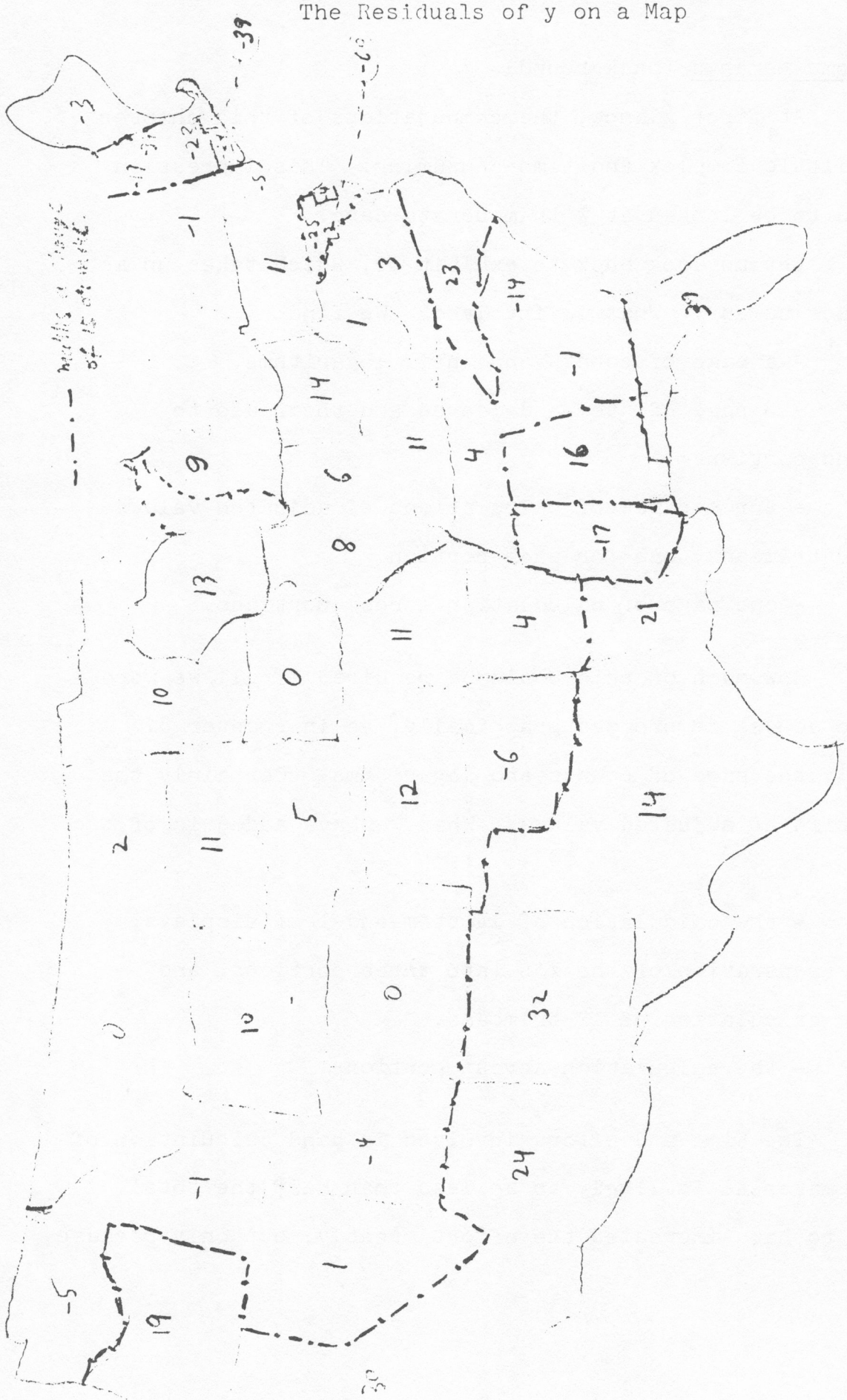
We would do violence to our principles if we did not look at our residuals on a map. Exhibit 15 shows the residuals, with state boundaries emphasized when the residual changes by more than 15. The result is a number of wholly enclosed patches, lying around the western, southern, and eastern edges of the country:

- Oregon to Louisiana, high.
- Mississippi, low.
- Alabama, high.
- Florida, higher.
- South Carolina, low.
- North Carolina, high.
- Maryland, low.
- D.C., lower.
- five New England states, low.

The appearance is clear. Detailed explanation may -- or may not -- be simple.

exhibit 15 of chapter 10

The Residuals of y on a Map



10G. Some remarks (background)

At first glance, the computations of this chapter seem a little complex and time-consuming. This impression deserves to be looked at with moderate care.

Let us look back to exhibit 12, which takes up as much space as any. What is involved? We find:

- a page of counts and their logarithms.
- a page of stem-and-leaves and their use to find portions
- three pages of calculations of adjusted values and trimeans, one for each portion.
- one page of calculation across portions.

How much of this would be required if all we were going to do was to proceed graphically, as in chapter 8? Certainly the page of counts and logarithms. Certainly the calculation of adjusted values. What we have added is of two sorts:

- the calculation of 14 stem-and-leaf displays, the separation of the x's into three portions, and the calculation of 12 trimeans.
- the calculation across portions.

The time and effort involved in hand calculation of the new material is likely to be less than half the total. We seem to have increased the effort greatly, but only because

we have put down the detailed arithmetic for calculations we would have had to make anyway. The actual increase is not that great.

And, as we feel now we have gone through it a few times, the complexity of the standard form for the calculation across portions is not so great, either.

***

10H. How far have we come?

This chapter has stressed the use of arithmetic to find tilts, most particularly for the purpose of seeing under first appearances by subtracting these tilts. The following points need to be constantly kept in mind:

- graphs are always competitive as ways of finding tilts, but may not win for a number of reasons, including:
  -- the tilt may be needed as a piece of a more complex calculation, ultimately to go on a computer.
  -- there may be too many points for an easily graspable figure.
  -- if things are simple enough -- a few points, equally spaced -- arithmetic can be faster (if the first step will meet our needs).

- the methods of this chapter work rather well, but they are not -- and should not -- be advertised as usually the best. We are explorers. We want to get there and back, safely.

- tilt is properly measured by finding what straight line to subtract - namely that which leaves what remains untilted.

These points are general, and not at all confined to the specific approach of this chapter.

We have chosen a simple pattern of approach:

- divide the data into three portions according to x.
- summarize each portion by a summary point.
- use the two end summary points to determine a tilt.
- subtract the tilt.
- using the same portions, resummarize and repeat as may be necessary.
- use all three summary points to determine a level.
- subtract this level to form residuals.
- take the total tilt subtracted in all steps and the level subtracted to specify the line fitted.

We have suggested and illustrated two of many ways to find summary points:

- find the means of x's and of y's --- no second step should be needed. Stray values can distort results.
- find middle means of x's and of y's or residuals --- repeated steps are likely to be needed, but the final answers are resistant to stray values.

Throughout we have continued to stress wise choice of expression, both of x and of y. At a purely practical level, we have chosen origin and unit, especially for x, to make arithmetic easy.

We have recognized that perfect untilting is never necessary for the effective examination of residuals. Accordingly, we have often only calculated the effect of a last step of untilting for the summary points and the fitted line, not bothering to adjust individual residuals.

*****

10K. A functional example (optional, but instructive)

Let us turn next to an example where irregularities are very small. Exhibit 16 gives 15 values of a certain mathematical function and proceeds to fit a straight line to them. The values of the function have been rounded from the nine decimals given in the source to five decimals. We have chosen to use trimeans, although the portions contain only five or six points. This is reasonable because the data is extremely smooth.

The fit found is, to six decimals

$$.398929 + .05139x$$

Note that we have tried to squeeze one more decimal out of the fit than we were given in the (rounded-down) data.

The residuals are plotted in exhibit 17, where an eye-fitted curve has been drawn in. The central points are also shown. The numerical values of the residuals are given in exhibit 18, together with values read off from the eye-fitted curve and the differences

residual - eye fit

Had our eye-fitted curve been a great success -- had we been completely successful in squeezing out the extra decimal, the eye-fitted curve would have passed through the points where each residual would have fallen had we used six-decimal values rather than five-decimal ones as data.

exhibit 16 of chapter 10

Fitting A Straight Line to a Certain Mathematical Function

A) DATA, PORTIONS, CENTRAL POINTS

x	y		
.000	.39894		
.004	.39914 H	x-trimean = .006	
.006	.39924 M	y-trimean = .39924	
.008	.39934 H		
.014	.39965		
.016	.39975		
.020	.39995		
.026	.40026		
.028	.40036		
.030	.40046		
.040	.40098		
.042	.40109 H	x-trimean = .044	
.044	.40119 M	y-trimean = .40120	
.046	.40130 H		
.050	.40151		

xxx

Portions

half of .000 to .050:	.000 to .024
	.026 to .050
third of count:	.000 to .014
	.040 to .050
results:	.000 to .014
	.040 to .050

(Cont'd)

exhibit 16 (cont'd.)

B) CALCULATION ACROSS PORTIONS

	for portions; basis means			(_ *)	apparent	trial
At x =	.006	.025	.044	(.038)		
Adjusted:						
raw y	.39924	.40021	.40120	(.00196)	.052	(.052)
1st adjusted	.398928		.398906	(-.000022)	-.00058	(-.00058)
2nd adjusted	.398932		.398931	(-.000001)	-.00003	(-.00003)
3rd adjusted	.398932	.398922	.398932	(.000006)	.00000	polish
same				[1.196782]	.398929	(.398929)
residual	.00003	-.000007	.000003	[-.000001]	.000000	polish
Fitted:						
1st tilt	.000312		.002288		.052x	
2nd tilt	-.000003		-.000026		-.00058x	
3rd tilt	-.000000	(.001285)	-.000004		-.00003x	
level	.398929	.398929	.398929			+ .398929
total	.399238	.400214	.401185		.05139x	+ .398929
(residual)	(.000003)	(-.000007)	(.000003)			
restored y	.399241	.400207	.401188			

Polished:
(none needed)

Translated:
$y \sim .05139x + .398929$

(*) In ( ), value for hi portion minus value for lo portion -- followed by tilts, divisor .038.

In [ ], sum of values for B portions -- followed by levels, divisor 3.

10-58

exhibit 16 (cont'd.)

C) LOW-x CALCULATIONS

x	y	.052x	Diff	-.00058x	Diff	-.00003x	Diff	.398929	Diff
.000	.39894	.0000	.398940	-.000000	.398940	-.000000	.398940	.398929	.000011
.004	.39914	.000208	.398932	-.000002	.398934	-.000000	.398934	.398929	.000005
.006	.39924	.000312	.398928	-.000003	.398931	-.000000	.398931	.398929	.000002
.008	.39934	.000416	.398924	-.000005	.398929	-.000000	.398929	.398929	.000000
.014	.39965	.000728	.398922	-.000008	.398930	-.000000	.398930	.398929	.000001
		H	.398924		.398930		.398930		.000001
		M	.398928		31		31		.000002
		H	.398932		34		34		.000005
		trimean	.398928		.398932		.398932		.000002

D) HIGH-x CALCULATIONS

.040	.40098	.002080	.398900	-.000023	.398923	-.000001	.398924	.398929	-.000005
.042	.40109	.002184	.398906	-.000024	.398930	-.000001	.398932	.398929	.000002
.044	.40119	.002288	.398902	-.000026	.398928	-.000001	.398929	.398929	.000000
.046	.40130	.002392	.398908	-.000027	.398935	-.000001	.398936	.398929	.000007
.050	.40151	.002600	.398910	-.000029	.398939	-.000002	.398941	.398929	.000012
		H	.398902		.398928		.398929		.000000
		M	.398906		30		32		.000002
		H	.398908		35		36		.000007
		trimean	.398906		.398931		.398932		.000003

10-59

exhibit 16 (cont'd.)

E) MID-PORTION CALCULATIONS

x	y	.05139x	Diff	.398929	Diff
.016	.39975	.000822	.398928	.398929	-.000001
.020	.39995	.001028	.398922	.398929	-.000007
.026	.40026	.001336	.398924	.398929	-.000005
.028	.40036	.001439	.398921	.398929	-.000008
.030	.40046	.001542	.398918	.398929	-.000011
H	.39995	.001028	.398921		?
M	.40026	.001336	22		?
H	.40036	.001439	24		
trimean	.40021	.001385	.398922		

Some tendency for the center to sag below the trimeans of .398932 of the end portions.

Overall adjustment: $\frac{1}{3}$ (.398932+.398922+.398932) = .398929

F) SOURCE

Table 9.8, at page 422, of Handbook of Mathematical Functions. National Bureau of Standards Applied Mathematics series 55. Washington, Government Printing Office, 1964.

exhibit 17 of chapter 10

The Residuals Eye-fitted

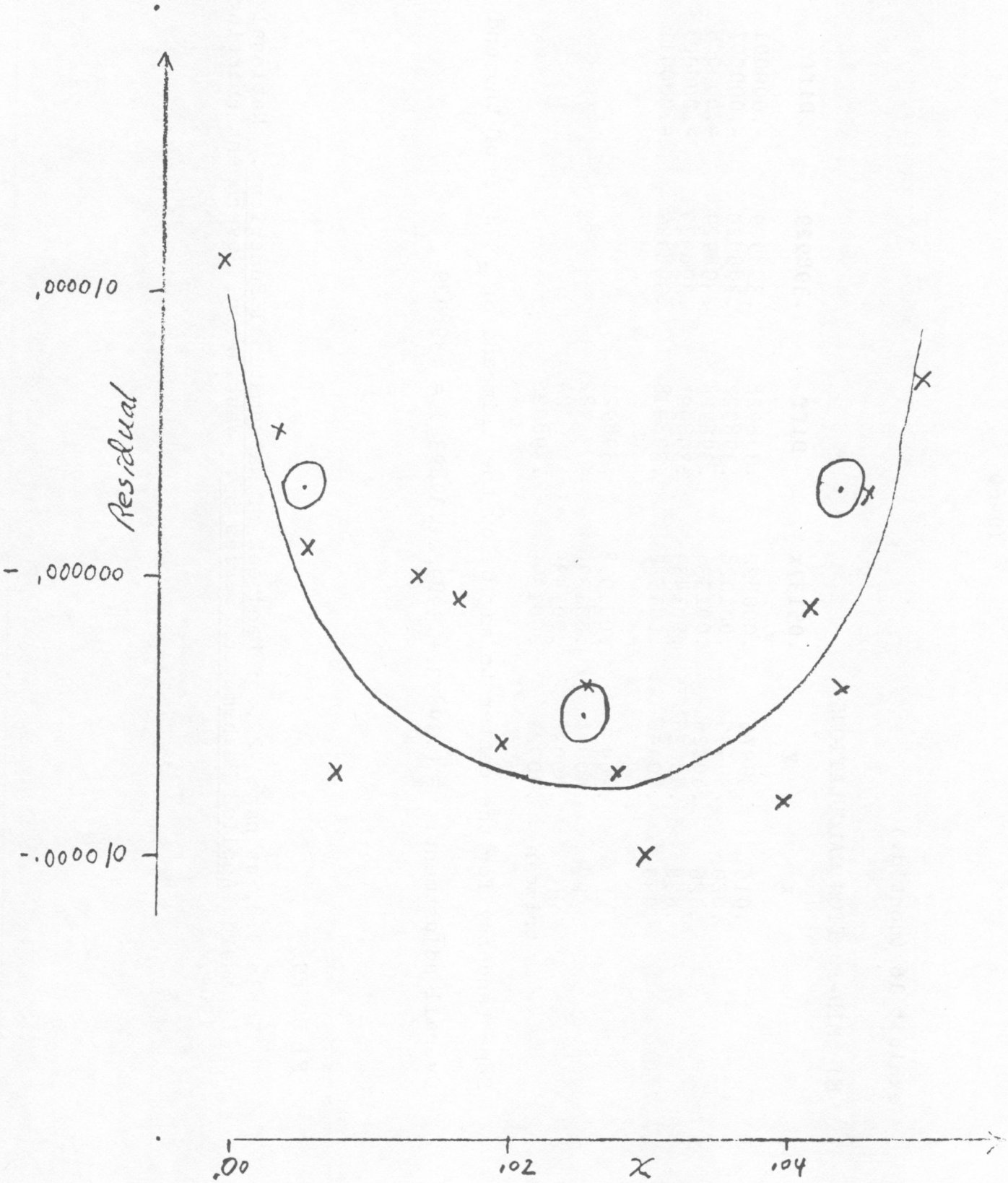


exhibit 18 of chapter 10

Did We Squeeze Out a 6th Decimal?

A) RESIDUALS, EYE-FIT, AND COMPARISON

x	residual	eye fit	diff	(*)
.000	.000011	.000012	.000001	-.000002
.004	.000005	.000006	.000001	-.000002
.006	.000002	.000003	.000001	-.000003
.008	.000000	.000001	.000001	-.000003
.014	.000001	-.000003	.000004	.000004
.016	-.000001	-.000004	.000003	.000003
.020	-.000007	-.000006	.000001	-.000001
.026	-.000005	-.000007	.000002	.000002
.028	-.000008	-.000006	.000002	-.000001
.030	-.000011	-.000006	-.000005	-.000004
.040	-.000005	.000001	.000006	-.000004
.042	.000002	.000003	.000001	+.000002
.044	.000000	.000005	.000005	-.000003
.046	.000007	.000006	.000001	.000002
.050	.000012	.000012	.000000	.000000

*5-decimal value less 6-decimal value, both from original table (as in panel F of exhibit 16). For x = .000 this involves .39894-.398942 = -.000002.

In this ideal state, then,

eye fit

and and

residual + (5-decimal value) - (6-decimal value)

would have been equal. This is equivalent to having the much smaller numbers

residual - eye fit = better residual

and and

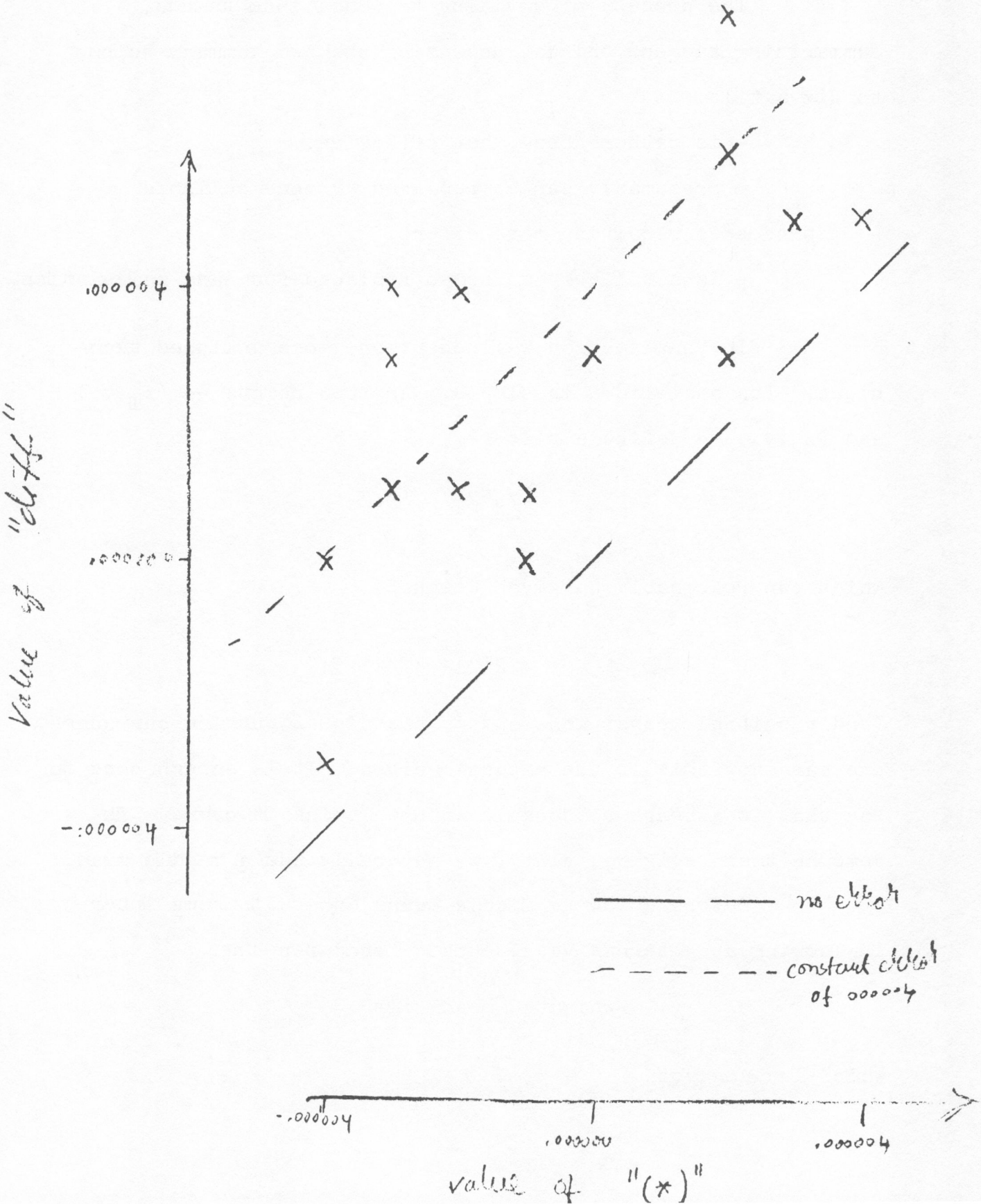
(*) = (6-decimal value) - (5-decimal value)

equal. Both are given in exhibit 18, numerically. The corresponding graph is shown as exhibit 19. We see that "diff" and "(*)" agree quite well -- only a few pairs differ by as much as .000002 from +.000002. This means that our eye-fitted curve has almost recovered the missing sixth decimal.

If we had a good way to fit a bend, and thus take the most of the droop out of exhibit 17, we might well do even better, coming even closer to squeezing out the additional ungiven decimal place. We can dream of doing this because our five-decimal values were rounded from a table that was correct to better than six decimals. The errors in the rounded values cannot exceed, say, ±5.1 in the sixth decimal, so that we do have a chance to squeeze out the sixth decimal. With most sets of data, such squeezing would be quite impractical.

exhibit 19 of chapter 10

Comparison of "Diff" With Its Hopeful Relation to "(*)"



*****

10L. <u>When computing is no pain</u> (optional background or optional with computer)

The process of dividing the data into thirds, summarizing the end thirds, and using the two summary points to fix a tilt:

- is rather crude, but effective.
- presumably can be replaced by more polished processes requiring more effort.
- is not likely to be so replaced for hand calculation.

In this section we look at one more polished technique. Its basic idea is simple. Any two points -- (x_1,y_1) and (x_2,y_2) -- define a tilt

$$\frac{y_2-y_1}{x_2-x_1}$$

which can reasonably be given weight

$$(x_2-x_1)^2 .$$

(Under suitable conditions and by techniques outside our scope, one can show this is the natural weight. It is enough here to say that it appears reasonable when we think it over.) To combine these weighted tilts, we naturally use a middle mean.

Before going to middle means, there is some interest in looking at alternative algebraic forms for the

weighted mean tilt

which include both

$$\frac{\frac{(y_1-y_2)}{(x_1-x_2)}(x_1-x_2)^2 + \frac{(y_1-y_3)}{(x_1-x_3)}(x_1-x_3)^2 + \ldots}{(x_1-x_2)^2 + (x_1-x_3)^2 + \ldots}$$

and

$$\frac{(y_1-y_2)(x_1-x_2) + (y_1-y_3)(x_1-x_3) + \ldots}{(x_1-x_2)^2 + (x_1-x_3)^2}$$

and

$$\frac{(y_1-\bar{y})(x_1-\bar{x}) + (y_2-\bar{y})(x_2-\bar{x}) + \ldots}{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots}$$

where the equality of the last two forms may not be obvious. The last of these is a form that has been used rather frequently.

Exhibit 20 shows this process in action in the specially simple case where the x's are equally spaced. We find that the slopes thus found fit in well with our earlier results, for we have the following tilts:

means of thirds:	63.2	(exhibit 5)
trimeans of thirds:	74.5	(exhibit 9)
weighted trimean:	79.8	
weighted midmean:	76.5	

exhibit 20 of chapter 10

The Weighted-Tilt Process Applied to the Data of exhibit 5

A) The ELEMENTARY TILTS, $(y_i-y_j)/(x_i-x_j)$ in TRIANGULAR FORM

873										
1374	1124									
-2769	-698	-174								
604	-1082	-264	20							
679	642	-495	-28	152						
-183	248	367	-417	-53	-96					
618	218	371	430	-210	54	171				
204	411	213	330	384	-141	75	175			
-246	-21	192	98	214	279	-156	35	128		
-94	-170	-45	120	60	163	226	-148	21	106	
-52	-73	-131	-47	86	41	132	191	-138	14	92
(Wts) (1)	(4)	(9)	(16)	(25)	(36)	(49)	(64)	(81)	(100)	(121)

B) ORDERED TILTS (WITH WEIGHTS)

1374(1)	367(9)	175(64)	75(49)	-47(16)	-174(9)
1124(4)	330(16)	171(49)	60(25)	-52(1)	-183(1)
873(1)	279(36)	163(36)	54(36)	-73(4)	-210(25)
679(1)	248(4)	153(25)	41(36)	-94(1)	-246(1)
642(4)	226(49)	132(49)	35(64)	-96(36)	-258(25)
618(1)	218(4)	128(81)	21(81)	-131(9)	-264(9)
604(1)	214(25)	120(16)	20(16)	-138(81)	-417(16)
430(16)	213(9)	106(100)	14(100)	-141(36)	-495(9)
411(4)	204(1)	98(16)	-21(4)	-148(64)	-698(4)
384(25)	192(9)	91(121)	-28(16)	-156(49)	-1082(4)
371(9)	191(64)	86(25)	-45(9)	-170(4)	-2769(1)
(Sum) (67)	(226)	(582)	(436)	(301)	(104) ✓

Sum of weights 67 + 226 + 582 + 436 + 301 + 104 = 1716

Now 1716 = 429 + 858 + 429 so that hinges are at 429h and 1287h, median at 858h.

C) HINGES, MEDIAN, AND TRIMEAN

Value	Ranks	(Inverse ranks)	
+163	1275 to 1310	(407 to 442)	U hinge
+92	831 to 951	(766 to 866)	median
-28	415 to 430	(1287 to 1302)	L hinge

$$\text{trimean} = \frac{163+92+92-28}{4} = 79.8$$

exhibit 20 (Cont'd)

D) MIDMEAN

$$(13)(163) + (25)(15\cdot) + (49)(132) + \ldots + (16)(20)$$

$$+ (81)(21) + (100)(14) + (4)(-21) + (1)(-28) = 65,665$$

$$(13) + (25) + (49) + \ldots \qquad + (81) + (169) = 858$$

$$\frac{65,665}{858} = 76.5$$

E) EXAMPLES (Wts)

$$\frac{3,904-3,031}{1919-1918} = 873 \qquad (1)$$

$$\frac{5,278-3,031}{1920-1918} = 1124 \qquad (4)$$

$$\frac{2,509-5,278}{1921-1920} = -2769 \qquad (1)$$

$$\frac{2,509-3,031}{1921-1918} = -174 \qquad (9)$$

* short-cutting? *

Since the number of pairs of observations rises rapidly as the number of observations increases, we would like to save calculation. It is natural to try to do this by selecting a few points and using only the tilts found by connecting each of these few points to each possible point. Unfortunately this does not seem to work as well as one would hope. Exhibit 21 shows this for a few choices and the familiar example. If we are to short cut, the methods of exhibits 3 and 5 and of exhibits 7 and 9 still look good, especially the latter. Exhibit 22 shows the results for our example. Rounding off the weights of exhibit 20 to one of a few equispaced values is a more reasonable process. The tilts of 61 and 64 thus obtained are moderately good, but probably not as good as using middle means of thirds.

Another approach is to use the weighted hinges of panel C of exhibit 20 to declare certain tilts as outside, and then to take the weighted mean of those that remain. Panel B of exhibit 23 shows the result. Panel D compares the various tilts based on means.

Another approach, useful for equally-spaced values is to combine elementary tilts of the same or similar weights by trimeans, accumulating weights, and then from the weighted midmean of the result. Exhibit 24 illustrates this procedure for our example.

exhibit 21 of chapter 10

Results of Using a Few Selected Points to Combine With All to Give Tilts Which Then Give a Weighted Midmean

Selected points	Tilt: trimean	Tilt: midmean
the mean of all 12	93	88
the midmean of 12	~100	~91
3 midmeans, one of each third	79	81
4 easy midmeans (= medians), one of each quarter	54	52

Other approaches	tilt
exhibit 5	63
exhibit 9	74
exhibit 21, trimean	72
exhibit 21, midmean	70

exhibit 22 of chapter 10

The tilts of exhibit 21 combined using only

3 weights other than 0

A) CONVERSION RULE USED for WEIGHTS

121,100,81,64	to 4
49,36	to 2
25,16	to 1
9,4,1	to 0

B) TILTS whose CONVERTED WEIGHTS are NON-ZERO

	430(1)	191(4)	128(4)	75(2)	20(1)	-141(2)
	384(1)	175(4)	121(1)	60(1)	14(4)	-148(4)
	330(1)	171(2)	106(4)	54(2)	-28(1)	-156(2)
	279(1)	163(2)	98(1)	41(1)	-47(1)	-210(1)
	226(2)	152(1)	92(4)	35(9)	-70(2)	-258(1)
	214(1)	132(2)	86(1)	21(4)	-138(4)	-417(1)
Sums	(7)	(15)	(15)	(19)	(13)	(11)

C) CALCULATION of WEIGHTED TRIMEAN and WEIGHTED MIDMEAN

80	
20h H	132
40h M	60
20h H	-7
tri	61

(1)(132) + (4)(128) + ... + (1)(20) + (4)(14)

= 2573

2573/40 = 64 = wtd. midmean

D) COMPARISON of TILTS based on MIDDLE-MEANS

exhibit 9	74
	80
exhibit 20, trimean	80
panel C, trimean	61
exhibit 20, midmean	76
panel C, midmean	64

exhibit 23 of chapter 10

Use of a skipping weighted-mean tilt

A) Which TILTS are OUTSIDE? -- hinges from panel C of exhibit 20

H	163	dh	191	S	354 (12 tilts left above, total wt = 76)
H	-28			S	-219 (8 tilts left below, total wt = 69)

B) REPETITION of panel A of exhibit 20, OMITTING OUTSIDE TILTS

	---	---									
	---	---	-174								
	---	---	---	20							
	---	---	---	-28	152						
	---	---	---	---	-59	-96					
	---	---	---	---	-210	54	171				
	---	248	---	---	---	-141	75	175			
	-183	218	---	330	214	279	-156	35	128		
	---	---	213	98	60	163	226	-148	21	106	
	204	-21	192	120	86	41	132	191	-138	14	92
	---	-170	-45	-47							
	-94	-73	-131								
	-52										
sum	-125	202	55	493	243	301	448	253	11	120	92
sum×wt	-125	808	495	3944	6075	10836	21952	16192	891	12000	11132 (84200)
sum of weights	4	20	45	96	150	216	245	256	243	200	121 (1596)

$$\text{skipping weighted-mean tilt} = \frac{84,200}{1,596} = 52.8$$

C) SIMILAR CALCULATION OMITTING NONE of the TILTS

$$\text{weighted mean tilt} = \frac{96,549}{1,716} = 56.3$$

(cont'd)

exhibit 23 (cont'd.)

D) SKIPPING APPLIED to exhibit 22 -- S points are -146 and 271

---	20	152	-96	171	175	128	106	92
---	-28	-59	54	75	35	21	14	
---	---	---	-141	---	---	-138		
---	---	---	---	226	191			
---	---	214	163	132				
---	98	60	41					
---	120	86						
---	-47							
#	10		9		9			
sum	616		625		624			
Wt	1		2		4			
sum × wt	616		1250		2496		(4362)	
# × wt	10		18		36		(64)	

skipping short-cut weighted tilt $\frac{4362}{64} = 68$

E) COMPARISON of TILTS based on MEANS

panel D 68
exhibit 5 63
panel C 56
panel B 53

exhibit 24 of chapter 10

Two-stage combination for equally-spaced x's

A) ELEMENTARY TILTS REARRANGED and their TRIMEANS -- see panel A of exhibit 20

	1374								
	873	1124					226		
	679H	612	371	430			191		
	618	411H	367	330h	384	279	175H	128	last 2
	604	248	213H	920h	214H	163H	171	106H	columns
	204M	218m	192	98m	152	54m	132M	92m	combine 2
	-52	-21m	-45M	20m	86M	41m	75	21m	and 3
	-94	-73	-131	-28h	60	-96H	35H	14H	levels of
	-183H	-178H	-174H	-47h	-59H	-111	-145	-138	weights,
	-246	-898	-264	-417	-210		-156		respect-
	-2769	-1082	-495						ively.
tri	226	104h	-13	76h	79	40h	118h	58	
(wt)	(11)	(40)	(81)	(128)	(175)	(216)	(501)	(564)	

B) WEIGHTED MIDMEAN of TRIMEAN TILT

	Trimean	**Wt**	**Middle Wt**	**Trimean × Middle Wt**
	226	11	-	-
	118h	501	83	9,798
	104h	40	40	4,180
	79	175	175	13,825
	76h	128	128	9,792
	58	564	432	25,056
	40h	216	-	-
	13	81	-	-
(sum)			(858)	(62,651)

$$\text{weighted midmean} = \frac{62,651}{858} = 73$$

10M <u>We could have skipped states</u> (optional, uses chapter 6)

In our analysis of which states seemed to require special attention so far as marriages and motor vehicle deaths are concerned, we used the hinges of all 49 residuals to identify which states were "outside". Of the nine states "outside" four -- D.C., R.I., N.H. and Fla. -- were "detached". It is certainly natural to set these 4 aside before determining which other states are "outside".

Panel A of exhibit 25 shows the result of doing just this. Vermont -- already spotted as a neighbor of N.H. and Mass. -- is now outside.

If we go on to a full skipping analysis -- as in Panel B of the same exhibit -- Mississippi and South Carolina come out low, while North Carolina and Arizona come out high.

Before leaving this example, we ought to ask about the consequences of skipping during the fitting. We leave it to the reader to show that using skipping trimeans instead of trimeans alters the fitted tilt by a moderate amount.

exhibit 25 of chapter 10

Skipping analyses of all 49 residuals

A) ANALYSIS with 4 INITIALLY DETACHED STATES SET ASIDE

3	-2	542	542	
6	-1	974	9	74
to 15	-0	655433111		655433111
45 ✓ 14	0	00011234456899		00011234456899'
to 16	1	01111234469		01111234469
5	2	134		134
2	3	02	02	
			(outside)	(inside)

45		dh	14
12h	-3	S	-17
12h	11	S	25

Additional outside state: Vermont (found as neighbor in exhibit 14).

B) FULL s-SKIPPING ANALYSIS

- - - - - - - - skipping calculations - - - - - - - - -

49	U	40	51	36	52	35	S3
13H	-4	17hH	-1	18hH	-h	18H	-1
13H	11	14hH	11	13hH	11	14H	11
dh	15	dh	12	dh	11h		stops
S	-19(6)	S	-13(9)	S	-12(9)		
S	26(3)	S	23(4)	S	22h(5)		

- - - -additional states skipping outside - - - - -

State	Residual	x	y
Vermont*	-19	-47	-11
Mississippi	-17	56	74
South Carolina	-14	59	79
North Carolina	23	43	103
Arizona*	24	0	71

*Also picked up in exhibit 14.

*****

10P. Problems for chapter 10

1. Consider the colorimetric data in problem 3 of chapter 8. Fit a tilt arithmetically. Plot the residuals. Discuss.

2. Consider the carbon content of 36 clays as given in exhibit 24 of chapter 8. Which clays fall in each of the three portions according to the simple rules of section 10A? Fit tilts to these values, using first means and then midmeans. Find and plot residuals.

3. Assume at least three values for the amount of slowly decaying radioactivity in the two left-hand portions of exhibit 22 of chapter 9. (For which manner of expressing amount of radioactivity?) Calculate the corresponding tilts using midmeans. Plot the residuals. Discuss.

4. Take the 1952 election results in exhibit 25 of chapter 8 as x and the 1956 results as y. Fit a tilt arithmetically. Find the residuals and plot them. Discuss the result.

5. Lowering of the freezing point of water by (ethyl) alcohol. (1,720) translates as: "For an alcohol concentration of 0.01 normal, the lowering of the freezing point in °C was 1.720 times the normality (hence it must have been 0.0172°C)." Another way to put this is that the freezing-point-lowering constant was 1.720. Data-sets (11): (1,1720), (2,1755), (3,1769),

(4,1822), (5,1787), (6,1799), (8,1816), (10,1875), (12,1826), (14,1832), (16,1829). Source: E. H. Loomis 1894. Ueber ein exacteres Vertahren bei die Bestimmung von Gefrierpunktserniedrigugen. Annalen der Physik und Chemie 51: 500-523. Table on page 517. Problem: Fit a tilt for the freezing-point-lowering constant against alcohol concentration arithmetically. Find residuals and plot them.

6. Vapor pressures of decahydro naphthalenes (two isomers). (-300,44) translates as: "For a temperature of -30.0C°, the vapor pressure in centimeters of mercury was 0.044." Data-sets for the trans isomer (24): (-300,44), (-241,60), (-111,107), (-07,173), (00,182), (130,287), (309,603), (389,823), (515,1202), (597,1672), (653,2000), (744,2702), (835,3818), (955,5501), (1124,9511), (1194,12110), (1367,20685), (1523,32065), (1680,48949), (1871,27605), (2088,128020), (2129,139640), (2234,172230), (2353,214530). Data-sets for the cis. isomer (23): (-295,38), (-192,65), (-100,086), (00,109), (120,148), (200,182), (380,344), (434,400), (499,530), (504,541), (529,606), (569,731), (600,824), (700,1257), (801,1950), (924,3281), (1051,5450), (1098,6315), (1124,6816), (1246,10521), (1487,22582), (1727,43428), (1947,76102). (The last translates as "At 194.7°C the vapor pressure was 76.102 centimeters of mercury". Source: W. F. Seyer and C. W. Mann 1945. The vapor pressures of cis- and trans-decahydro naphthalene. J. Amer. Chem. Soc. 67: 328-329. Table 1 on page 328. Problem: Re-express the variables

properly and fit a tilt for the (re-expressed) vapor pressure of one isomer against the (re-expressed) temperature. Find the residuals. Discuss the result.

7. Repeat 6 for the other isomer. Compare and contrast the results.

8. Select some other batch of (x,y) data from chapters 8, 9 or 11 and fit a straight line arithmetically. (Any respective fit has to include plotting residuals.)

9. Find, from other sources, a batch of data of interest to you that includes at least 20 data sets, and fit a line arithmetically. (See remark closing problem 8.)